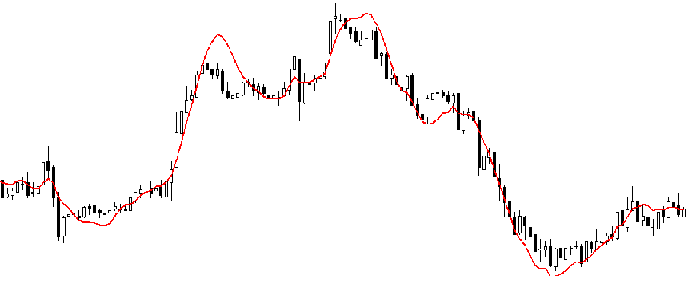
1. **Define Regression. Explain the concepts of regression and types in detail. (14 marks)**

What is Regression Analysis?

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent**(target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the [causal effect relationship](https://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/) between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Regression_Line.png)

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

Why do we use Regression Analysis?

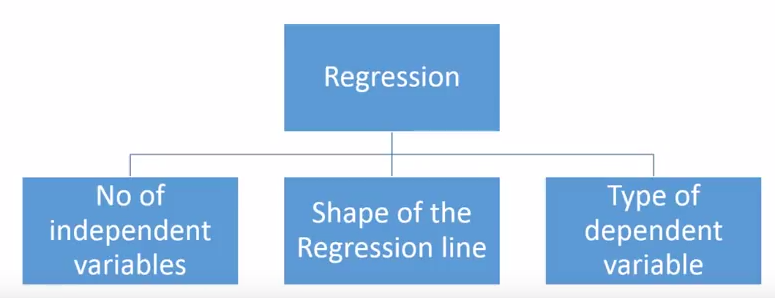
There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

How many types of regression techniques do we have?

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics-number of independent variables, type of dependent variables and shape of regression line.

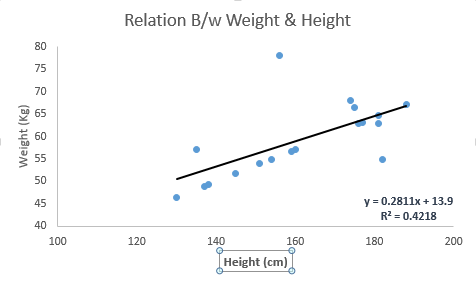
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Regression_Type.png)

1. Linear Regression

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

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

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

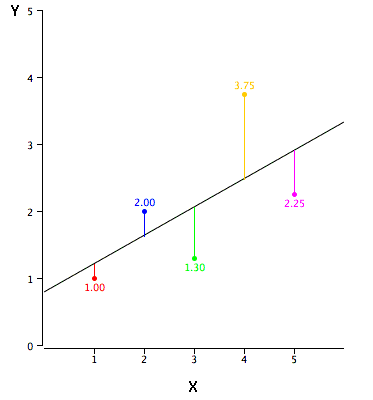
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Linear_Regression1.png)

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

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

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

[least square, regression line](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Least_Square.png)

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/reg_error.gif)

We can evaluate the model performance using the metric **R-square**. Important Points:

* There must be **linear relationship** between independent and dependent variables
* Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
* Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
* Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to
* minor changes in the model. The result is that the coefficient estimates are unstable
* In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.

2. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

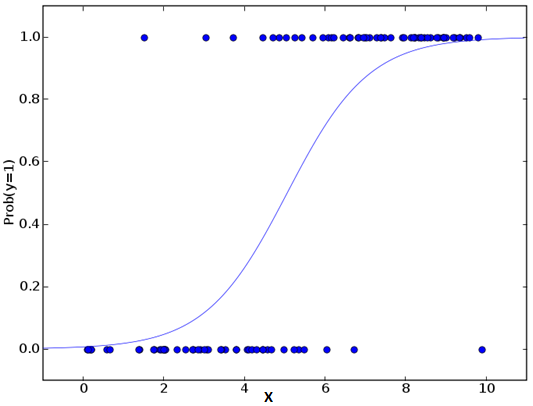
odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk+

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is “why have we used log in the equation?”.

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is logit function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Logistic_Regression.png)

Important Points:

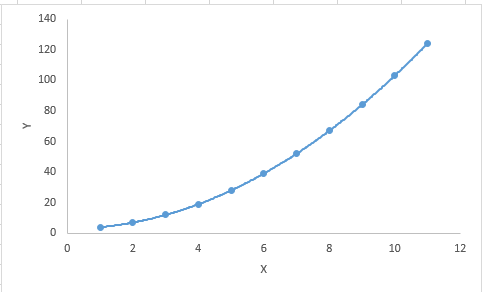
* Logistic regression is widely used for **classification problems**
* Logistic regression doesn’t require linear relationship between dependent and independent variables.  It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio
* To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression
* It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square
* The independent variables should not be correlated with each other i.e. **no multicollinearity**.  However, we have the options to include interaction effects of categorical variables in the analysis and in the model.
* If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**
* If dependent variable is multi class then it is known as **Multinomial Logistic regression**.

3. Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

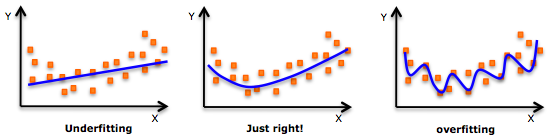
y=a+b\*x^2

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Polynomial.png)

Important Points:

* While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/underfitting-overfitting.png)

* Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation.

4. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below:

* Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
* Forward selection starts with most significant predictor in the model and adds variable for each step.
* Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

5. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Above, we saw the equation for linear regression. Remember? It can be represented as:

y=a+ b\*x

This equation also has an error term. The complete equation becomes:

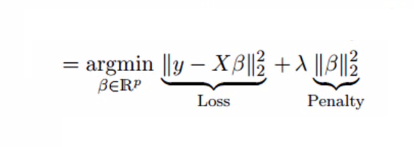
y=a+ b\*x+ e (error term),

[Error term is the value needed to correct for a prediction error between the observed and predicted value]

=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

In a linear equation, prediction errors can be decomposed into two sub components. First is due to the **biased** and second is due to the **variance**. Prediction error can occur due to any one of these two or both components. Here, we’ll discuss about the error caused due to variance.

Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda). Look at the equation below.

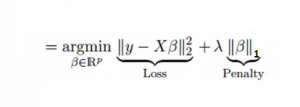
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Ridge2.png)

In this equation, we have two components. First one is least square term and other one is lambda of the summation of β2 (beta- square) where β is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

Important Points:

* The assumptions of this regression is same as least squared regression except normality is not to be assumed
* Ridge regression shrinks the value of coefficients but doesn’t reaches zero, which suggests no feature selection.
* This is a regularization method and uses l2 regularization.

6. Lasso Regression

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Lasso.png)

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models.  Look at the equation below: Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.

Important Points:

* The assumptions of lasso regression is same as least squared regression except normality is not to be assumed.
* Lasso Regression shrinks coefficients to zero (exactly zero), which certainly helps in feature selection.
* Lasso is a regularization method and uses l1 regularization.
* If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero.

7. ElasticNet Regression

ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

[elastic net regression](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Elastic_Net.png)

A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge’s stability under rotation.

Important Points:

* It encourages group effect in case of highly correlated variables
* There are no limitations on the number of selected variables
* It can suffer with double shrinkage

Beyond these 7 most commonly used regression techniques, you can also look at other models like Bayesian, Ecological and Robust regression.

How to select the right regression model?

Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:

1. First step in selecting the right model is to identify the relationship and impact of variables.
2. To compare the goodness of fit for different models, analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term, and check for possible bias in the model, by comparing the model with all possible submodels.
3. Cross-validation is the best way to evaluate models used for prediction. Divide the data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give a measure for the prediction accuracy.
4. It’ll also depend on the objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.
5. Regression regularization methods (Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.

**2.Explain Linear Regression with example.(14 marks)**

**Introduction**

Models use machine learning algorithms, during which the machine learns from the data just like humans learn from their experiences. Machine learning models can be broadly divided into two categories based on the learning algorithm which can further be classified based on the task performed and the nature of the output.

1.**Supervised learning methods:** It contains past data with labels which are then used for building the model.

* **Regression**: The output variable to be predicted is *continuous*in nature, e.g. scores of a student, diamond prices, etc.
* **Classification**: The output variable to be predicted is *categorical*in nature, e.g. classifying incoming emails as spam or ham, Yes or No, True or False, 0 or 1.

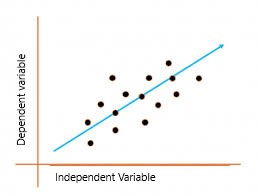
2. **Unsupervised learning methods:** It contains no predefined labels assigned to the past data.

* **Clustering**: No predefined labels are assigned to groups/clusters formed, e.g. customer segmentation.

Linear Regression is a supervised learning algorithm in machine learning that supports finding the *linear*correlation among variables. The result or output of the regression problem is a real or continuous value.

Simple Linear Regression

Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression*.*If there is a single input variable **X**(independent variable), such linear regression is called ***simple linear regression***.



The above graph presents the linear relationship between the output(y) variable and predictor(X) variables.  The blue line is referred to as the*best fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

*T*o calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

**Yi= β0 + β1Xi**

where Yi = Dependent variable,  **β0** = constant/Intercept, **β1** = Slope, **Xi** = Independent variable.

**Linear Regression Algorithm**

This algorithm explains the linear relationship between the dependent (output) variable y and the independent (predictor) variable X using a straight line  Y= B0 + B1 X.

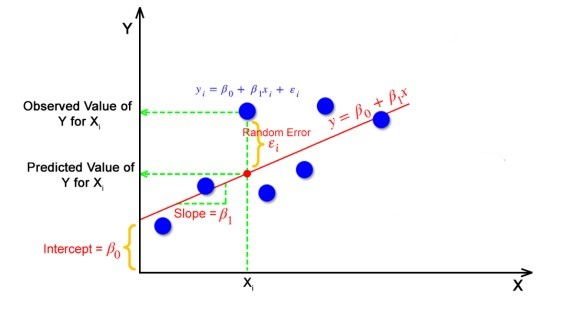
**In simple terms:**It is a method of finding the best straight line fitting to the given dataset, i.e. tries to find the best linear relationship between the independent and dependent variables.

It is mostly done with the help of the **Sum of Squared Residuals Method**, known as the **Ordinary least squares (OLS) method**.

## Why do we square the residuals instead of using modulus?

The squared function is differentiable everywhere, while the absolute error is not differentiable at all the points in its domain (its derivative is undefined at 0). This makes the squared error more preferable to the techniques of mathematical optimization. To optimize the squared error, we can compute the derivative and set its expression equal to 0, and solve. But to optimize the absolute error, we require more complex techniques having more computations.

Actually, we use the Root Mean Squared Error instead of Mean squared error so that the unit of RMSE and the dependent variable are equal and results are interpretable.



But how the linear regression finds out which is the best fit line?

The goal of the linear regression algorithm is to get the **best values for B0 and B1** to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

**Random Error(Residuals)**

In regression, the difference between the observed value of the dependent variable(**yi**) and the predicted value(**predicted**) is called the residuals.

**εi=** **ypredicted** –   **yi**

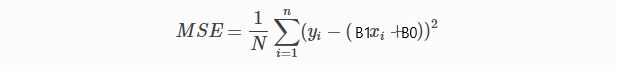
**where ypredicted =   B0 + B1 Xi**

**Cost Function for Linear Regression**

The Cost Function helps to work out the optimal values for B0 and B1, which provides the best fit line for the data points.

In Linear Regression, generally **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the **ypredicted** and **yi**.

We calculate MSE using simple linear equation y= mx + b:

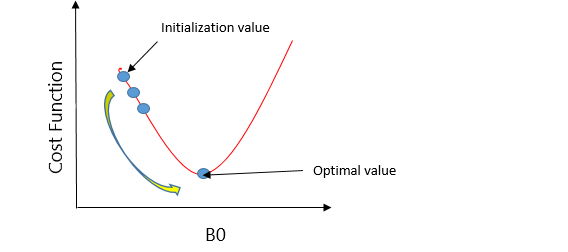


Using the MSE function, we will update the values of B0 and B1 such that the MSE value settles at the minima.  These parameters can be determined using the gradient descent method such that the value for the cost function is minimum.

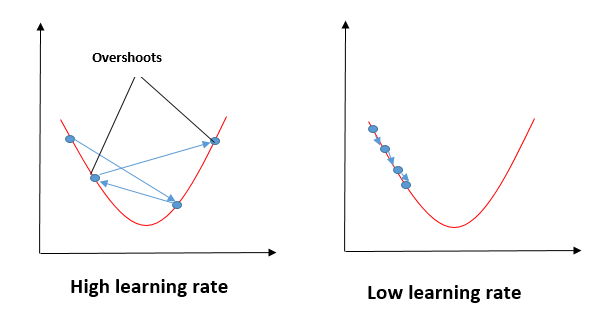
**Gradient Descent for Linear Regression**

Gradient Descent is one of the optimization algorithms that optimize the cost function (objective function) to reach the optimal minimal solution. To find the optimum solution we need to reduce the cost function (MSE) for all data points. This is done by updating the values of B0 and B1 iteratively until we get an optimal solution.

A regression model optimizes the gradient descent algorithm to update the coefficients of the line by reducing the cost function by randomly selecting coefficient values and then iteratively updating the values to reach the minimum cost function.

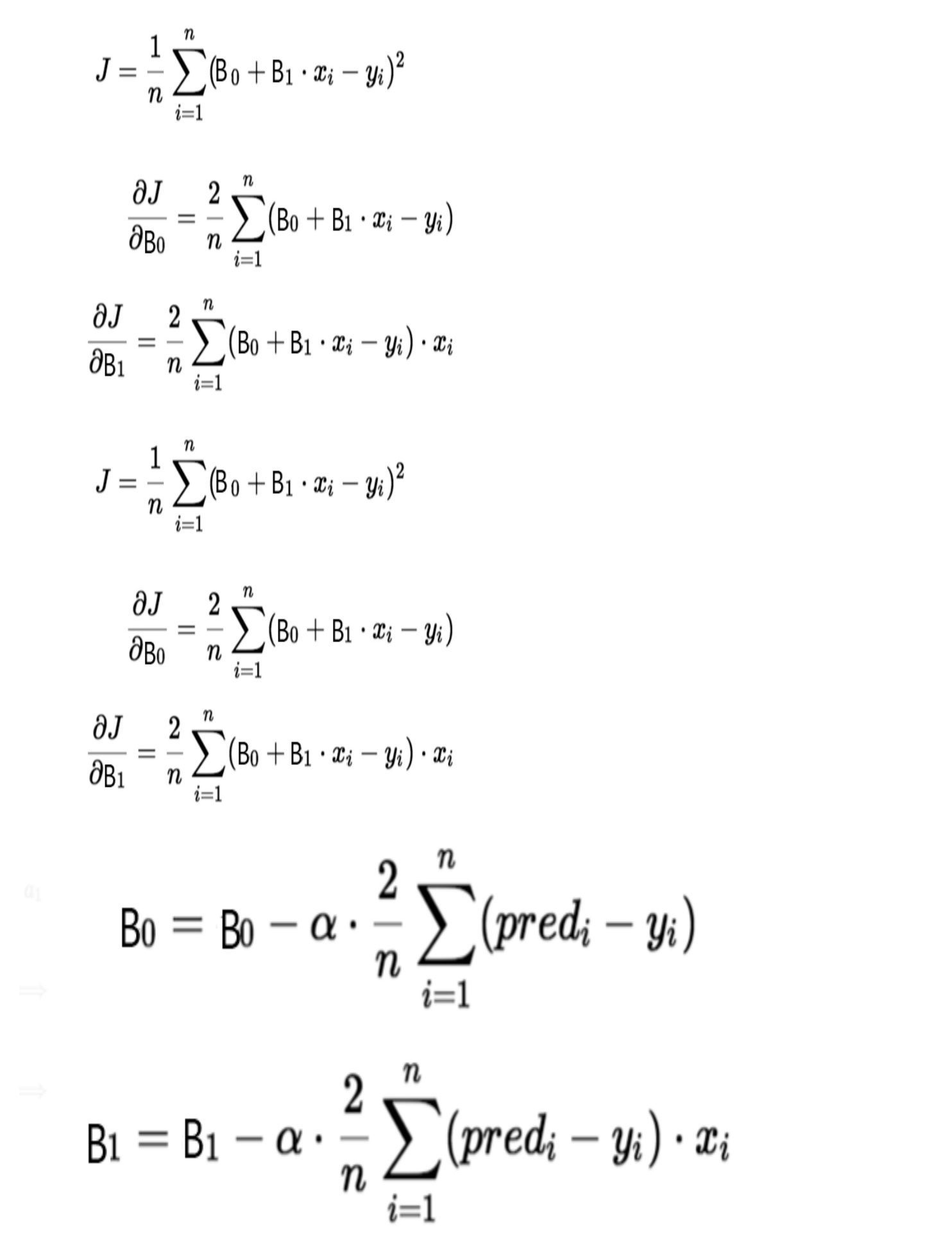


Let’s take an example to understand this. Imagine a U-shaped pit. And you are standing at the uppermost point in the pit, and your motive is to reach the bottom of the pit. Suppose there is a treasure at the bottom of the pit, and you can only take a discrete number of steps to reach the bottom. If you opted to take one step at a time, you would get to the bottom of the pit in the end but, this would take a longer time. If you decide to take larger steps each time, you may achieve the bottom sooner but, there’s a probability that you could overshoot the bottom of the pit and not even near the bottom. In the gradient descent algorithm, the number of steps you’re taking can be considered as the **learning rate**, and this decides how fast the algorithm **converges** to the minima.



To update B0 and B1, we take gradients from the cost function. To find these gradients, we take partial derivatives for B0 and B1.

We need to minimize the cost function J. One of the ways to achieve this is to apply the batch gradient descent algorithm. In batch gradient descent, the values are updated in each iteration. ( Below Last two equations shows the updating of values)

Alpha is the learning rate.

**Evaluation Metrics for Linear Regression**

The strength of any linear regression model can be assessed using various evaluation metrics. These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are,

1. Coefficient of Determination or R-Squared (R2)
2. Root Mean Squared Error (RSME) and Residual Standard Error (RSE)

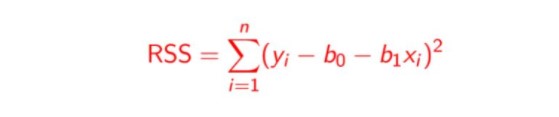
**1. Coefficient of Determination or R-Squared (R2)**

R-Squared is a number that explains the amount of variation that is explained/captured by the developed model. It always ranges between 0 & 1 . Overall, the higher the value of R-squared, the better the model fits the data.

Mathematically it can be represented as,

**R2 = 1 – ( RSS/TSS )**

* **Residual sum of Squares (RSS)** is defined as the sum of squares of the residual for each data point in the plot/data. It is the measure of the difference between the expected and the actual observed output.

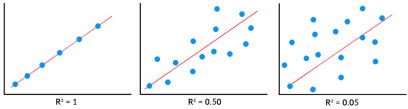


* **Total Sum of Squares (TSS)** is defined as the sum of errors of the data points from the mean of the response variable. Mathematically TSS is,

Total Sum of Squares

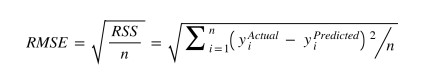
where y hat is the mean of the sample data points.

The significance of R-squared is shown by the following figures,

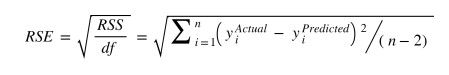


**2. Root Mean Squared Error**

The Root Mean Squared Error is the square root of the variance of the residuals. It specifies the absolute fit of the model to the data i.e. how close the observed data points are to the predicted values. Mathematically it can be represented as,



To make this estimate unbiased, one has to divide the sum of the squared residuals by the **degrees of freedom** rather than the total number of data points in the model. This term is then called the **Residual Standard Error(RSE)**. Mathematically it can be represented as,

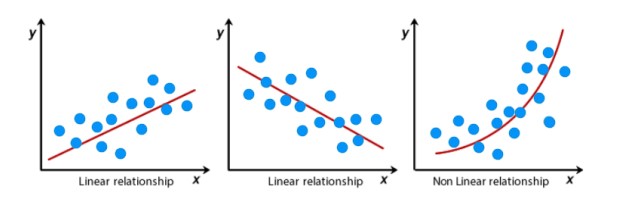


R-squared is a better measure than RSME. Because the value of Root Mean Squared Error depends on the units of the variables (i.e. it is not a normalized measure), it can change with the change in the unit of the variables.

**Assumptions of Linear Regression**

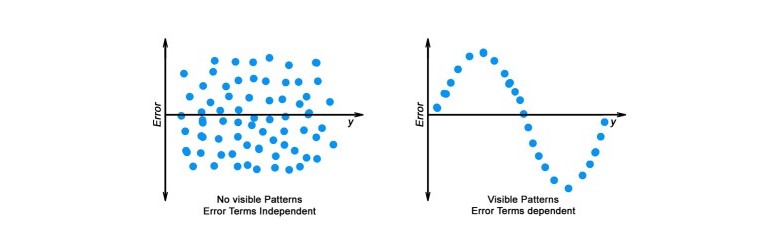
Regression is a parametric approach, which means that it makes assumptions about the data for the purpose of analysis. For successful regression analysis, it’s essential to validate the following assumptions.

1. **Linearity of residuals**: There needs to be a linear relationship between the dependent variable and independent variable(s).

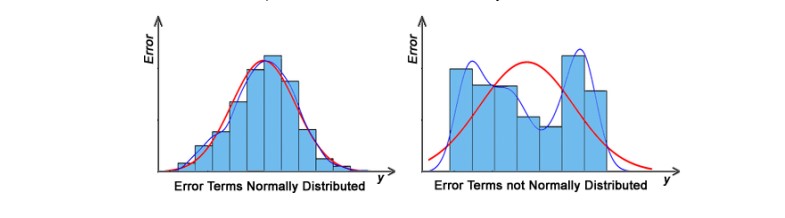


*2.***Independence of residuals:**The error terms should not be dependent on one another (like in time-series data wherein the next value is dependent on the previous one). There should be no correlation between the residual terms. The absence of this phenomenon is known as **Autocorrelation.**

There should not be any visible patterns in the error terms.

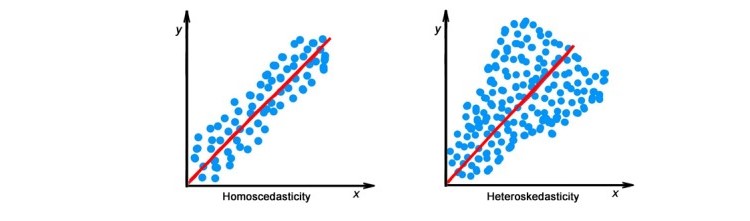


**3. Normal distribution of residuals:** The mean of residuals should follow a normal distribution with a mean equal to zero or close to zero. This is done in order to check whether the selected line is actually the line of best fit or not.  
If the error terms are non-normally distributed, suggests that there are a few unusual data points that must be studied closely to make a better model.



**4. The equal variance of residuals:** The error terms must have constant variance. This phenomenon is known as **Homoscedasticity**.

The presence of non-constant variance in the error terms is referred to as **Heteroscedasticity**. Generally, non-constant variance arises in the presence of *outliers or extreme leverage values*.



**Hypothesis in Linear Regression**

Once you have fitted a straight line on the data, you need to ask, “**Is this straight line a significant fit for the data?**” Or “I**s the** **beta coefficient explain the variance in the data plotted?”**And here comes the idea of **hypothesis testing** on the beta coefficient.

The Null and Alternate hypotheses in this case are:

H0: B1 = 0

HA: B1 **≠** 0

To test this hypothesis we use a **t-test,**test statistics for the beta coefficient is given by,

**Assessing the model fit(model fit statistics)**

Some other parameters to assess a model are:

1. **t statistic:** It is used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. **F statistic**: It is used to assess whether the overall model fit is significant or not. Generally, the higher the value of the F-statistic, the more significant a model turns out to be.

Multiple Linear Regression

Multiple linear regression is a technique to understand the relationship between a *single*dependent variable and *multiple*independent variables.

The formulation for multiple linear regression is also similar to simple linear regression with the small change that instead of having one beta variable, you will now have betas for all the variables used. The formula is given as:

 Y = B0 + B1X1 + B2X2 + … + BpXp + **ε**

Considerations of Multiple Linear Regression

All the four assumptions made for Simple Linear Regression still hold true for Multiple Linear Regression along with a few new additional assumptions.

1. **Over fitting**: When more and more variables are added to a model, the model may become far too complex and usually ends up memorizing all the data points in the training set. This phenomenon is known as the overfitting of a model. This usually leads to high training accuracy and very low test accuracy.
2. **Multicollinearity**: It is the phenomenon where a model with several independent variables, may have some variables interrelated.
3. **Feature Selection:** With more variables present, selecting the optimal set of predictors from the pool of given features (many of which might be redundant) becomes an important task for building a relevant and better model.

**Multicollinearity**

As multicollinearity makes it difficult to find out which variable is actually contributing towards the prediction of the response variable, it leads one to conclude incorrectly, the effects of a variable on the target variable.  Though it does not affect the precision of the predictions, it is essential to properly detect and deal with the multicollinearity present in the model, as random removal of any of these correlated variables from the model causes the coefficient values to swing wildly and even change signs.

Multicollinearity can be detected using the following methods.

1. **Pairwise Correlations:** Checking the pairwise correlations between different pairs of independent variables can throw useful insights in detecting multicollinearity.
2. **Variance Inflation Factor (VIF):** Pairwise correlations may not always be useful as it is possible that just one variable might not be able to completely explain some other variable but some of the variables combined could be ready to do this. Thus, to check these sorts of relations between variables, one can use VIF. VIF basically explains the relationship of one independent variable with all the other independent variables. VIF is given by,

multicollinearity VIF

*where i* refers to the*ith* variable which is being represented as a linear combination of the rest of the independent variables.

The common heuristic followed for the VIF values is if VIF > 10 then the value is definitely high and it should be dropped. And if the VIF=5 then it may be valid but should be inspected first. If VIF < 5, then it is considered a good vif value.

Overfitting and Underfitting in Linear Regression

There have always been situations where a model performs well on training data but not on the test data. While training models on a dataset, overfitting, and underfitting are the most common problems faced by people.

Before understanding overfitting and underfitting one must know about bias and variance.

**Bias:**

Bias is a measure to determine how accurate the model is likely to be on future unseen data. Complex models, assuming there is enough training data available, can do predictions accurately. Whereas the models that are too naive, are very likely to perform badly with respect to predictions. Simply, Bias is errors made by training data.

Generally, linear algorithms have a high bias which makes them fast to learn and easier to understand but in general, are less flexible. Implying lower predictive performance on complex problems that fail to meet the expected outcomes.

**Variance:**

Variance is the sensitivity of the model towards training data, that is it quantifies how much the model will react when input data is changed.

Ideally, the model shouldn’t change too much from one training dataset to the next training data, which will mean that the algorithm is good at picking out the hidden underlying patterns between the inputs and the output variables.

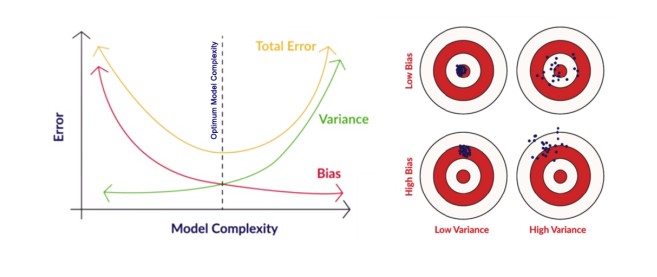
Ideally, a model should have lower variance which means that the model doesn’t change drastically after changing the training data(it is generalizable). Having higher variance will make a model change drastically even on a small change in the training dataset.

Let’s understand what is a bias-variance tradeoff is.

**Bias Variance Tradeoff**

The aim of any supervised machine learning algorithm is to achieve low bias and low variance as it is more robust. So that the algorithm should achieve better performance.

There is no escape from the relationship between bias and variance in machine learning.



There is an inverse relationship between bias and variance,

* An increase in bias will decrease the variance.
* An increase in the variance will decrease the bias.

There is a trade-off that plays between these two concepts and the algorithms must find a balance between bias and variance.

As a matter of fact, one cannot calculate the real bias and variance error terms because we do not know the actual underlying target function.

Now coming to the overfitting and underfitting.

Overfitting

When a model learns each and every pattern and noise in the data to such extent that it affects the performance of the model on the unseen future dataset, it is referred to as ***overfitting***. The model fits the data so well that it interprets noise as patterns in the data.

When a model has low bias and higher variance it ends up memorizing the data and causing overfitting. Overfitting causes the model to become specific rather than generic. This usually leads to high training accuracy and very low test accuracy.

Detecting overfitting is useful, but it doesn’t solve the actual problem. There are several ways to prevent overfitting, which are stated below:

* Cross-validation
* If the training data is too small to train add more relevant and clean data.
* If the training data is too large, do some feature selection and remove unnecessary features.
* Regularization

Underfitting:

When the model fails to learn from the training dataset and is also not able to generalize the test dataset, is referred to as ***underfitting***. This type of problem can be very easily detected by the performance metrics.

When a model has high bias and low variance it ends up not generalizing the data and causing underfitting. It is unable to find the hidden underlying patterns from the data. This usually leads to low training accuracy and very low test accuracy. The ways to prevent underfitting are stated below,

* Increase the model complexity
* Increase the number of features in the training data
* Remove noise from the data.

**Explain Least Square Estimation with example.**

Least Square Method Definition

The least-squares method is a crucial statistical method that is practised to find a regression line or a best-fit line for the given pattern. This method is described by an equation with specific parameters. The method of least squares is generously used in evaluation and regression. In regression analysis, this method is said to be a standard approach for the approximation of sets of equations having more equations than the number of unknowns.

The method of least squares actually defines the solution for the minimization of the sum of squares of deviations or the errors in the result of each equation. Find the [formula for sum of squares of errors](https://byjus.com/regression-sum-of-squares-formula/), which help to find the variation in observed data.

The least-squares method is often applied in data fitting. The best fit result is assumed to reduce the sum of squared errors or residuals which are stated to be the differences between the observed or experimental value and corresponding fitted value given in the model.

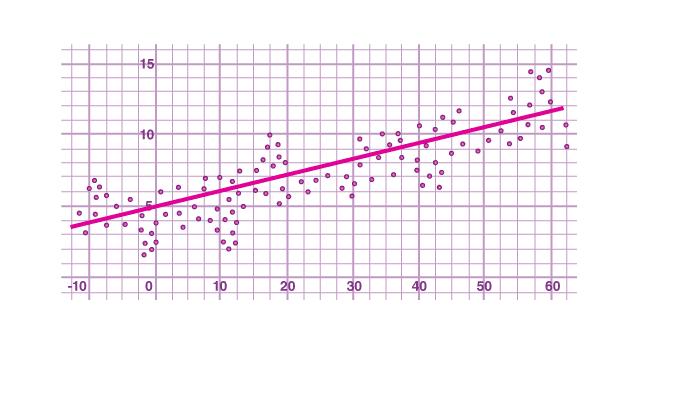
There are two basic categories of least-squares problems:

* Ordinary or linear least squares
* Nonlinear least squares

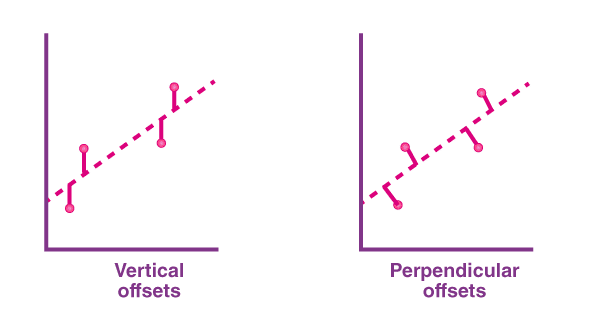
These depend upon linearity or nonlinearity of the residuals. The linear problems are often seen in regression analysis in statistics. On the other hand, the non-linear problems are generally used in the iterative method of refinement in which the model is approximated to the linear one with each iteration.

Least Square Method Graph

In linear regression, the line of best fit is a straight line as shown in the following diagram:



The given data points are to be minimized by the method of reducing residuals or offsets of each point from the line. The vertical offsets are generally used in surface, polynomial and hyperplane problems, while perpendicular offsets are utilized in common practice.



### **Least Square Method Formula**

The least-square method states that the curve that best fits a given set of observations, is said to be a curve having a minimum sum of the squared residuals (or deviations or errors) from the given data points. Let us assume that the given points of data are (x1, y1), (x2, y2), (x3, y3), …, (xn, yn) in which all x’s are independent variables, while all y’s are dependent ones. Also, suppose that f(x) is the fitting curve and d represents error or deviation from each given point.

Now, we can write:

d1= y1− f(x1)

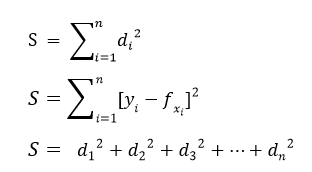
d2= y2− f(x2)

d3= y3− f(x3)

…..

dn= yn – f(xn)

The least-squares explain that the curve that best fits is represented by the property that the sum of squares of all the deviations from given values must be minimum, i.e:



Sum = Minimum Quantity

Suppose when we have to determine the equation of line of best fit for the given data, then we first use the following formula.

The equation of least square line is given by Y = a + bX

Normal equation for ‘a’:

∑Y = na + b∑X

Normal equation for ‘b’:

∑XY = a∑X + b∑X2

Solving these two normal equations we can get the required trend line equation.

Thus, we can get the line of best fit with formula y = ax + b

### **Solved Example**

The Least Squares Model for a set of data (x1, y1), (x2, y2), (x3, y3), …, (xn, yn) passes through the point (xa, ya) where xa is the average of the xi‘s and ya is the average of the yi‘s. The below example explains how to find the equation of a straight line or a least square line using the least square method.

**Question:**

Consider the time series data given below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| xi | 8 | 3 | 2 | 10 | 11 | 3 | 6 | 5 | 6 | 8 |
| yi | 4 | 12 | 1 | 12 | 9 | 4 | 9 | 6 | 1 | 14 |

Use the least square method to determine the equation of line of best fit for the data. Then plot the line.

**Solution:**

Mean of xi values = (8 + 3 + 2 + 10 + 11 + 3 + 6 + 5 + 6 + 8)/10 = 62/10 = 6.2

Mean of yi values = (4 + 12 + 1 + 12 + 9 + 4 + 9 + 6 + 1 + 14)/10 = 72/10 = 7.2

Straight line equation is y = a + bx.

The normal equations are

∑y = an + b∑x

∑xy = a∑x + b∑x2

|  |  |  |  |
| --- | --- | --- | --- |
| x | y | x2 | Xy |
| 8 | 4 | 64 | 32 |
| 3 | 12 | 9 | 36 |
| 2 | 1 | 4 | 2 |
| 10 | 12 | 100 | 120 |
| 11 | 9 | 121 | 99 |
| 3 | 4 | 9 | 12 |
| 6 | 9 | 36 | 54 |
| 5 | 6 | 25 | 30 |
| 6 | 1 | 36 | 6 |
| 8 | 14 | 64 | 112 |
| ∑x = 62 | ∑y = 72 | ∑x2 = 468 | ∑xy = 503 |

Substituting these values in the normal equations,

10a + 62b = 72….(1)

62a + 468b = 503….(2)

(1) × 62 – (2) × 10,

620a + 3844b – (620a + 4680b) = 4464 – 5030

-836b = -566

b = 566/836

b = 283/418

b = 0.677

Substituting b = 0.677 in equation (1),

10a + 62(0.677) = 72

10a + 41.974 = 72

10a = 72 – 41.974

10a = 30.026

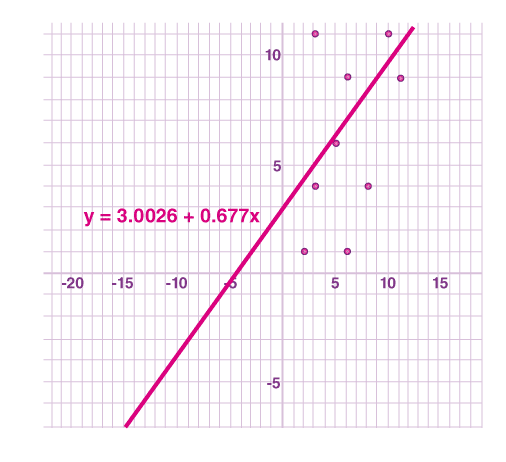
a = 30.026/10

a = 3.0026

Therefore, the equation becomes,

y = a + bx

y = 3.0026 + 0.677x



This is the required trend line equation.

Now, we can find the sum of squares of deviations from the obtained values as:

d1 = [4 – (3.0026 + 0.677\*8)] = (-4.4186)

d2 = [12 – (3.0026 + 0.677\*3)] = (6.9664)

d3 = [1 – (3.0026 + 0.677\*2)] = (-3.3566)

d4 = [12 – (3.0026 + 0.677\*10)] = (2.2274)

d5 = [9 – (3.0026 + 0.677\*11)] =(-1.4496)

d6  = [4 – (3.0026 + 0.677\*3)] = (-1.0336)

d7 = [9 – (3.0026 + 0.677\*6)] = (1.9354)

d8 = [6 – (3.0026 + 0.677\*5)] = (-0.3876)

d9 = [1 – (3.0026 + 0.677\*6)] = (-6.0646)

d10 = [14 – (3.0026 + 0.677\*8)] = (5.5814)

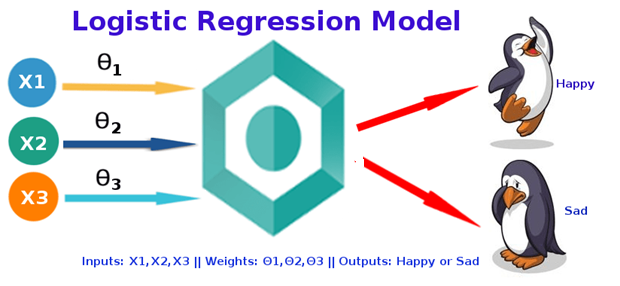
∑d2 = (-4.4186)2 + (6.9664)2 + (-3.3566)2 + (2.2274)2 + (-1.4496)2 + (-1.0336)2 + (1.9354)2 + (-0.3876)2 + (-6.0646)2 + (5.5814)2 = 159.27990

### **Limitations for Least-Square Method**

The least-squares method is a very beneficial method of curve fitting. Despite many benefits, it has a few shortcomings too. One of the main limitations is discussed here.

In the process of regression analysis, which utilizes the least-square method for curve fitting, it is inevitably assumed that the errors in the independent variable are negligible or zero. In such cases, when independent variable errors are non-negligible, the models are subjected to measurement errors. Therefore, here, the least square method may even lead to hypothesis testing, where parameter estimates and confidence intervals are taken into consideration due to the presence of errors occurring in the independent variables.

**Explain Logistic Regression Model Theory.**



* + How Linear regression is similar to logistic regression?
  + Derivation of the sigmoid function
  + What are odds?

4) Cost function in Logistic regression

5) What is the use of MLE in logistic regression?

* + Derivation of the Cost function
  + Why do we take the Negative log-likelihood function?

6) Gradient Descent Optimization

* Derivative of the Cost function
* Derivative of the sigmoid function

7) Endnotes

What is Logistic Regression?

**Logistic regression** is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

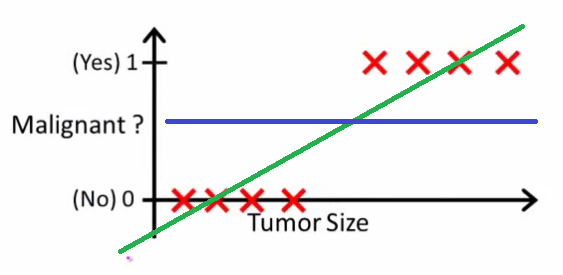
A person will survive this accident or not, The student will pass this exam or not. The outcome can either be yes or no (2 outputs). This regression technique is similar to linear regression and can be used to predict the **Probabilities** for classification problems.

Why do we use Logistic Regression rather than Linear Regression?

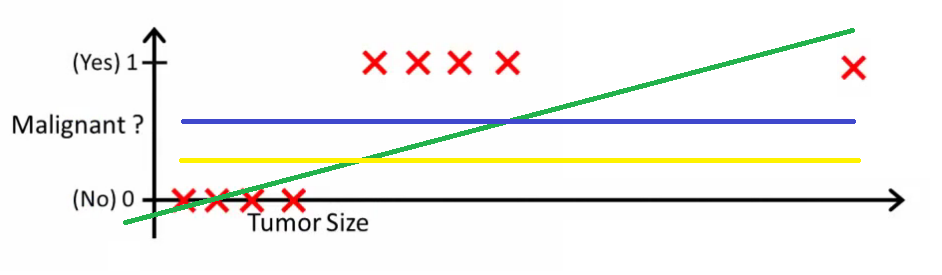
If you have this doubt, then you’re in the right place, my friend. After reading the definition of logistic regression we now know that it is only used when our dependent variable is binary and in linear regression this dependent variable is continuous.

The second problem is that if we add an outlier in our dataset, the best fit line in linear regression shifts to fit that point.

Now, if we use linear regression to find the best fit line which aims at minimizing the distance between the predicted value and actual value, the line will be like this:



Here the threshold value is 0.5, which means if the value of h(x) is greater than 0.5 then we predict malignant tumor (1) and if it is less than 0.5 then we predict benign tumor (0). Everything seems okay here but now let’s change it a bit, we add some outliers in our dataset, now this best fit line will shift to that point. Hence the line will be somewhat like this:

Do you see any problem here? The blue line represents the old threshold and the yellow line represents the new threshold which is maybe 0.2 here. To keep our predictions right we had to lower our threshold value. Hence we can say that linear regression is prone to outliers. Now here if h(x) is greater than 0.2 then only this regression will give correct outputs.

Another problem with linear regression is that the predicted values may be out of range. We know that probability can be between 0 and 1, but if we use linear regression this probability may exceed 1 or go below 0.

To overcome these problems we use Logistic Regression, which converts this straight best fit line in linear regression to an S-curve using the sigmoid function, which will always give values between 0 and 1. How does this work and what’s the math behind this will be covered in a later section?

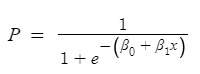
If you want to know the difference between logistic regression and linear regression then you refer to this [article](https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/).

Logistic Function

You must be wondering how logistic regression squeezes the output of linear regression between 0 and 1. If you haven’t read my [article](https://www.analyticsvidhya.com/blog/2021/07/practical-applications-of-linear-regression-models/) on Linear Regression then please have a look at it for a better understanding.

 Well, there’s a little bit of math included behind this and it is pretty interesting trust me.

Let’s start by mentioning the formula of logistic function:

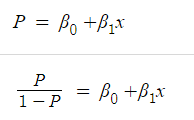


How similar it is too linear regression? If you haven’t read my article on Linear Regression, then please have a look at it for a better understanding.

 We all know the equation of the best fit line in linear regression is:

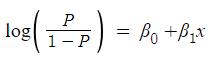
fit linear regression

Let’s say instead of y we are taking probabilities (P). But there is an issue here, the value of (P) will exceed 1 or go below 0 and we know that range of Probability is (0-1). To overcome this issue we take ***“odds”*** of P:

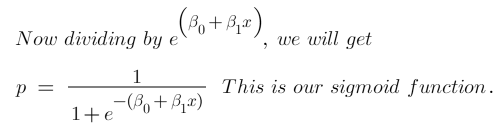
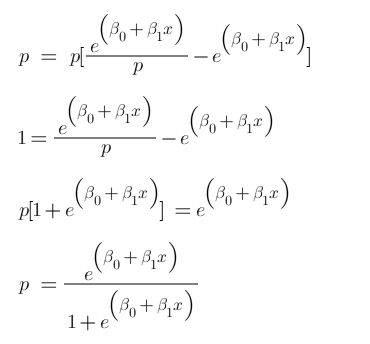
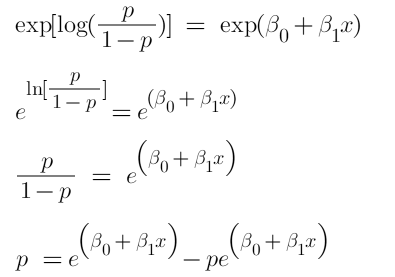


Do you think we are done here? No, we are not. We know that odds can always be positive which means the range will always be (0,+∞ ). Odds are nothing but the ratio of the probability of success and probability of failure. Now the question comes out of so many other options to transform this why did we only take ***‘odds’***? Because odds are probably the easiest way to do this, that’s it.

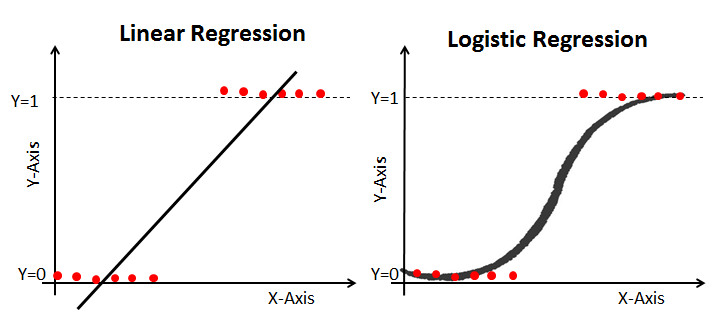
The problem here is that the range is restricted and we don’t want a restricted range because if we do so then our correlation will decrease. By restricting the range we are actually decreasing the number of data points and of course, if we decrease our data points, our correlation will decrease. It is difficult to model a variable that has a restricted range. To control this we take the ***log of odds***which has a range from (-∞,+∞).



If you understood what I did here then you have done 80% of the maths. Now we just want a function of P because we want to predict probability right? not log of odds. To do so we will multiply by ***exponent*** on both sides and then solve for P.

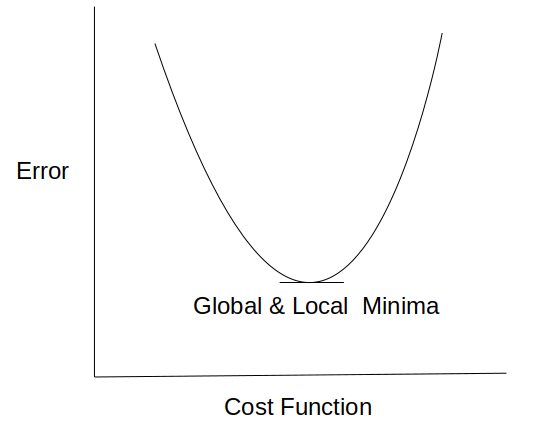


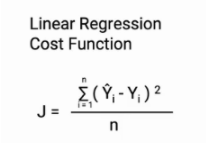
Now we have our logistic function, also called a sigmoid function. The graph of a sigmoid function is as shown below. It squeezes a straight line into an S-curve.



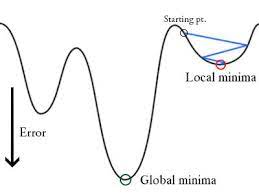
Cost Function in Logistic Regression

In linear regression, we use the Mean squared error which was the difference between y\_predicted and y\_actual and this is [derived](https://youtu.be/2PfGO753UHk) from the maximum likelihood estimator. The graph of the cost function in linear regression is like this:

**Image**

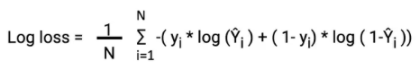


In logistic regression Yi is a non-linear function (*Ŷ*=1​/1+ e-z). If we use this in the above MSE equation then it will give a non-convex graph with many local minima as shown



The problem here is that this cost function will give results with local minima, which is a big problem because then we’ll miss out on our global minima and our error will increase.

In order to solve this problem, we derive a different cost function for logistic regression called ***log loss*** which is also derived from the *maximum likelihood estimation* method.



In the next section, we’ll talk a little bit about the maximum likelihood estimator and what it is used for. We’ll also try to see the math behind this log loss function.

What is the use of Maximum Likelihood Estimator?

The main aim of MLE is to find the value of our parameters for which the likelihood function is *maximized*. The likelihood function is nothing but a joint pdf of our sample observations and joint distribution is the multiplication of the *conditional probability* for observing each example given the distribution parameters. In other words, we try to find such that plugging these estimates into the model for P(x), yields a number close to one for people who had a malignant tumor and close to 0 for people who had a benign tumor.

Let’s start by defining our likelihood function. We now know that the labels are binary which means they can be either yes/no or pass/fail etc. We can also say we have two outcomes success and failure. This means we can interpret each label as Bernoulli random variable.

A random experiment whose outcomes are of two types, success S and failure F, occurring with probabilities p and q respectively is called a Bernoulli trial. If for this experiment a random variable X is defined such that it takes value 1 when S occurs and 0 if F occurs, then X follows a Bernoulli Distribution.

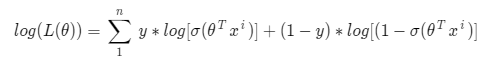
sigmoid**Where P is our sigmoid function**

sigmoid

where **σ(**θ^T\*x^i**)**is the sigmoid function. Now for n observations,

n observations

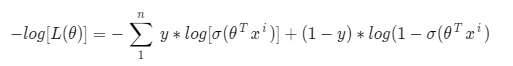
We need a value for theta which will maximize this likelihood function. To make our calculations easier we multiply the log on both sides. The function we get is also called the log-likelihood function or sum of the log conditional probability



In machine learning, it is conventional to minimize a loss(error) function via gradient descent, rather than maximize an objective function via gradient ascent. If we maximize this above function then we’ll have to deal with gradient ascent to avoid this we take negative of this log so that we use gradient descent. We’ll talk more about gradient descent in a later section and then you’ll have more clarity. Also, remember,

***max[log(x)] = min[-log(x)]***

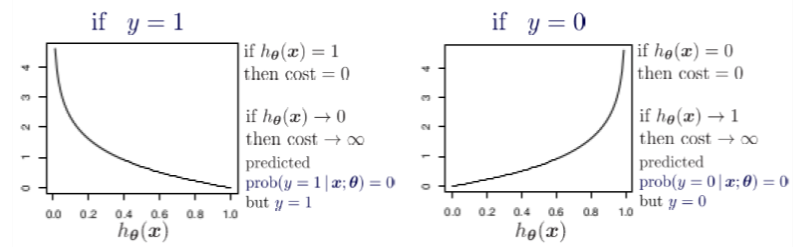
The negative of this function is our ***cost function*** and what do we want with our cost function? That it should have a minimum value. It is common practice to minimize a cost function for optimization problems; therefore, we can invert the function so that we minimize the negative log-likelihood (NLL). So in logistic regression, our cost function is:

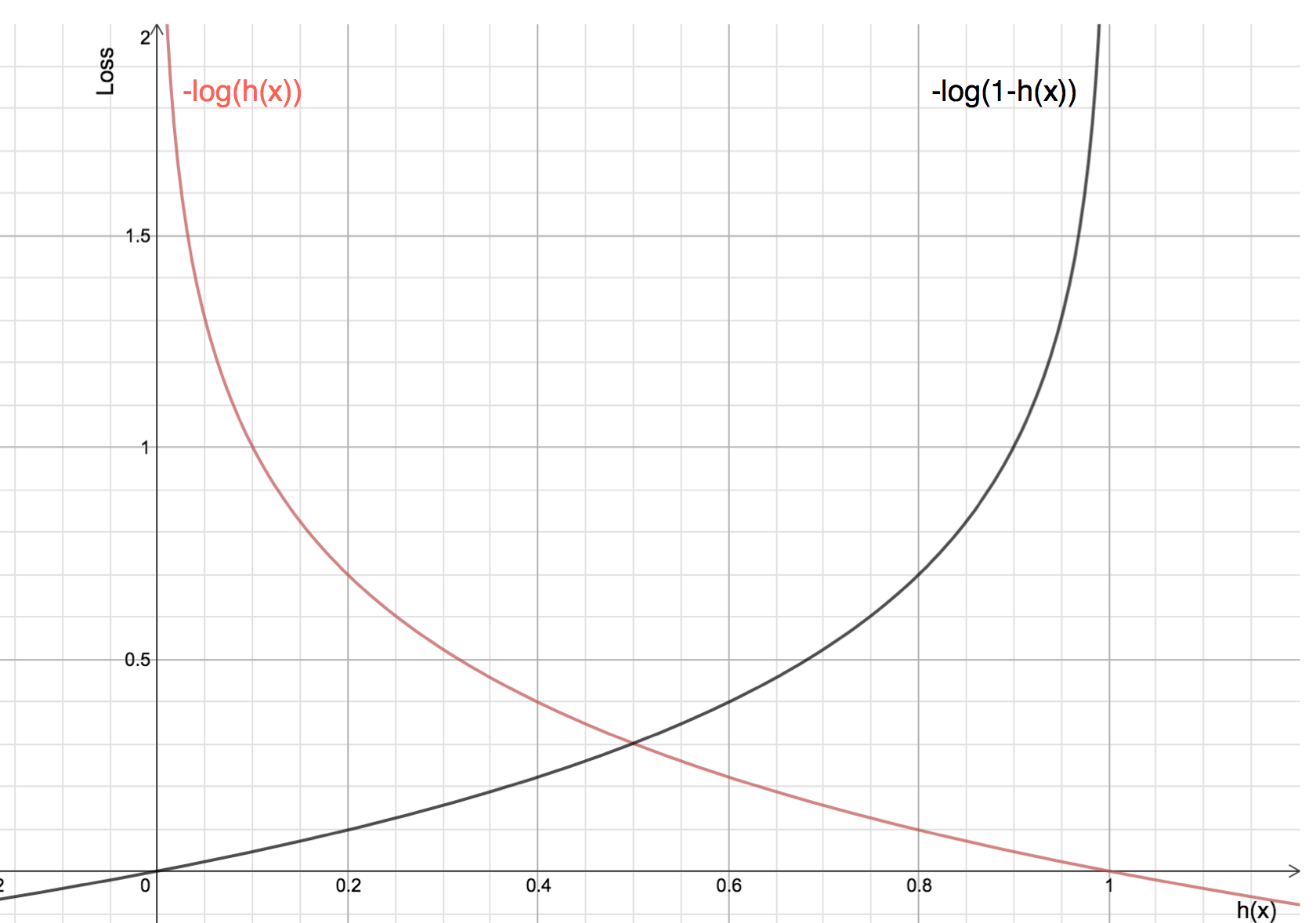


Here y represents the actual class and log(**σ(**θ^T\*x^i**) )**is the probability of that class.

* p(y) is the probability of 1.
* 1-p(y) is the probability of 0.

Let’s see what will be the graph of cost function when y=1 and y=0

If we combine both the graphs, we will get a convex graph with only 1 local minimum and now it’ll be easy to use gradient descent here.

The red line here represents the 1 class (y=1), the right term of cost function will vanish. Now if the predicted probability is close to 1 then our loss will be less and when probability approaches 0, our loss function reaches infinity.

The black line represents 0 class (y=0), the left term will vanish in our cost function and if the predicted probability is close to 0 then our loss function will be less but if our probability approaches 1 then our loss function reaches infinity.

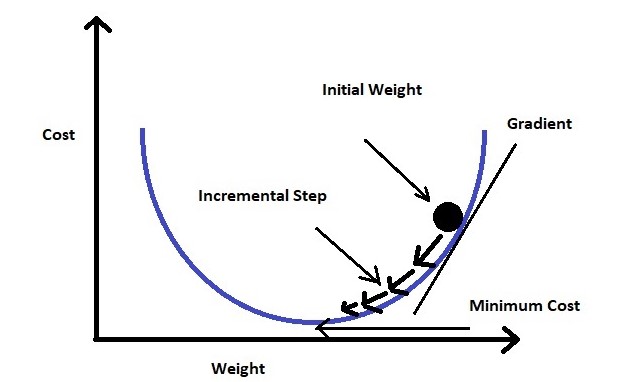
cost theta

This cost function is also called log loss. It also ensures that as the probability of the correct answer is maximized, the probability of the incorrect answer is minimized. Lower the value of this cost function higher will be the accuracy.

Gradient Descent Optimization

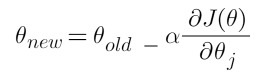
In this section, we will try to understand how we can utilize *Gradient Descent* to compute the minimum cost.

Gradient descent changes the value of our weights in such a way that it always converges to minimum point or we can also say that, it aims at finding the optimal weights which minimize the loss function of our model. It is an iterative method that finds the minimum of a function by figuring out the slope at a random point and then moving in the opposite direction.

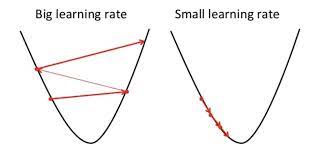
The intuition is that if you are hiking in a canyon and trying to descend most quickly down to the river at the bottom, you might look around yourself 360 degrees, find the direction where the ground is sloping the steepest, and walk downhill in that direction.

At first gradient descent takes a random value of our parameters from our function. Now we need an algorithm that will tell us whether at the next iteration we should move left or right to reach the minimum point. The gradient descent algorithm finds the slope of the loss function at that particular point and then in the next iteration, it moves in the opposite direction to reach the minima. Since we have a convex graph now we don’t need to worry about local minima. A convex curve will always have only 1 minima.

We can summarize the gradient descent algorithm as:



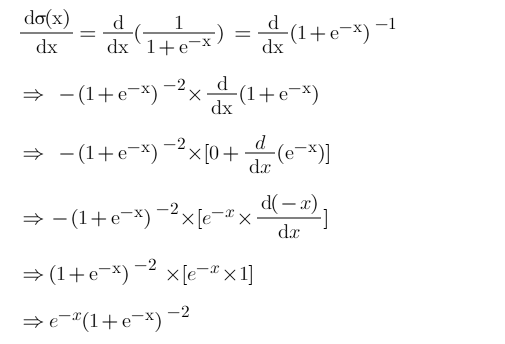
Here ***alpha***is known as the learning rate. It determines the step size at each iteration while moving towards the minimum point. Usually, a lower value of ***“alpha”***is preferred, because if the learning rate is a big number then we may miss the minimum point and keep on oscillating in the convex curve

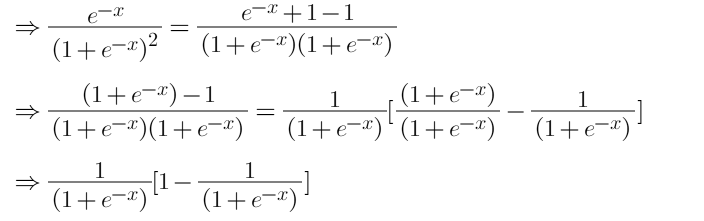


Now the question is what is this derivative of cost function? How do we do this? Don’t worry, In the next section we’ll see how we can derive this cost function w.r.t our parameters.

Derivation of Cost Function:

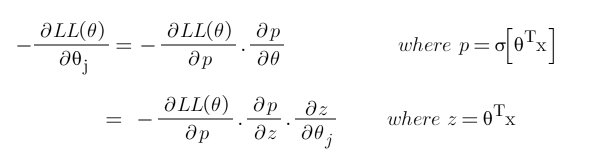
Before we derive our cost function we’ll first find a derivative for our sigmoid function because it will be used in derivating the cost function.



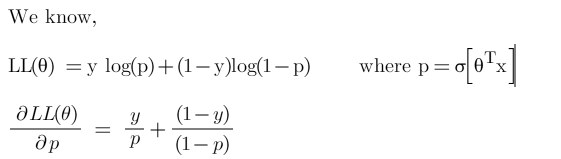


Now, we will derive the cost function with the help of the chain rule as it allows us to calculate complex partial derivatives by breaking them down.

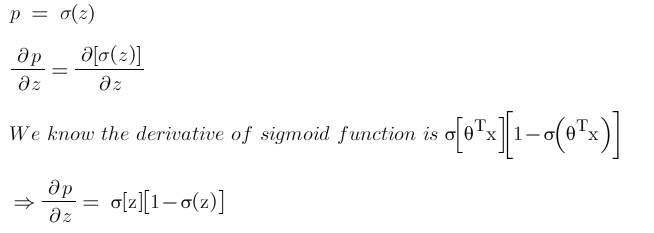
Step-1: Use chain rule and break the partial derivative of log-likelihood.



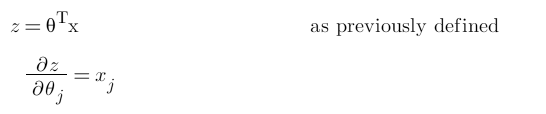
Step-2: Find derivative of log-likelihood w.r.t p



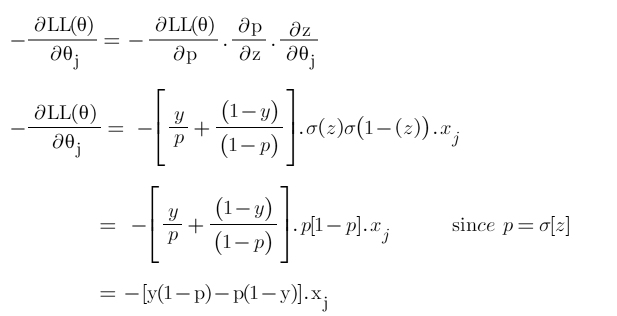
Step-3: Find derivative of ***‘p’*** w.r.t ***‘z’***

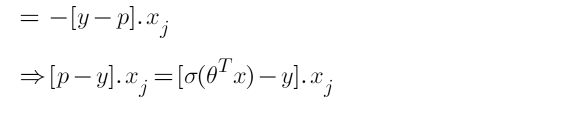


***Step-4: Find derivate of z w.r.t***θ

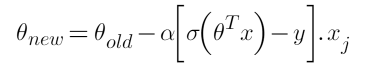


Step-5: Put all the derivatives in equation 1





Hence the derivative of our cost function is:



Now since we have our derivative of the cost function, we can write our gradient descent algorithm as:

If the slope is negative (downward slope) then our gradient descent will add some value to our new value of the parameter directing it towards the minimum point of the convex curve. Whereas if the slope is positive (upward slope) the gradient descent will minus some value to direct it towards the minimum point.

Endnote

To summarise, in this article we learned why linear regression doesn’t work in the case of classification problems.  Also, how MLE is used in logistic regression and how our cost function is derived.

**Explain Analytics applications to various Business Domains.**

With an interpretation of the business issue and becoming familiar with the accessible dataset, the business analyst gives insights into the growth of the business , also the companies can get ready to overcome the upcoming challenges. By these techniques, most problems will be solved already before the issue arises.

****

*Business Analytics Applications*

**8 Applications of Business Analytics**

Here, an index of industry application is taken into account for showing the role of business analyst in different domains.

1. Agriculture Business Analytics
2. Stock Marketing
3. Finance Marketing
4. Manufacturing Industry
5. Medical Methodology
6. Customer Relation Management
7. Bond Marketing
8. Human Resources

**1. Agricultural Business Analytics**

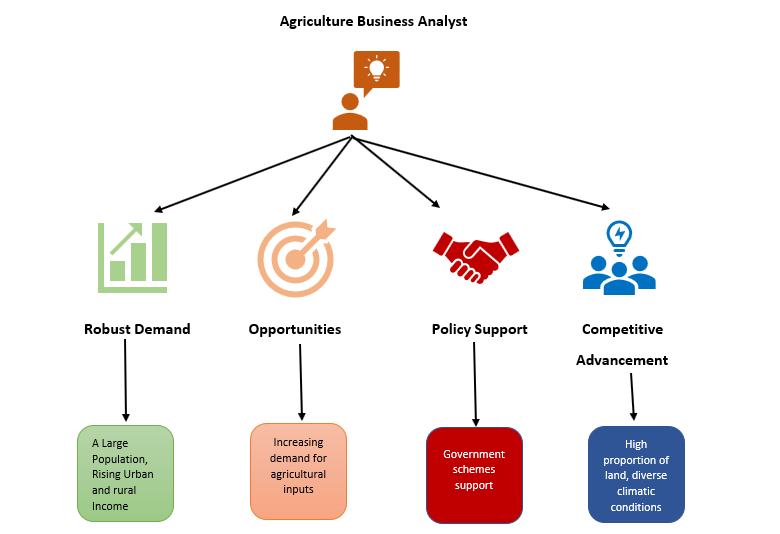
A large part of the Indian economy depends on agriculture. This sector has a major contribution to the economic growth of the country. However, You must know the [difference between economic growth with economic development](https://www.analyticssteps.com/blogs/difference-between-economic-growth-and-economic-development).

Indian agriculture experiences severe and drastic climatic conditions like depletion of ground-level water in rural areas, climate change, the emigration of farmhands from rural areas to urban areas in search of employment, urbanization, etc.

So, agriculture needs a lot of attention to focus on better investment and growth.

A business analyst can assure the availability of crops on time, crop production, quality and quantity of seeds, the effect of climate change, monsoon changes, rainwater storage, crop loss, fertilizer requirements, wind direction, floods, and draughts risk management, etc. can be controlled by predictions of business analysts.

For example, predictions on soil parameters can be analysed by previous data and implemented on crop yield.

****

*Agriculture Business Analytics*

[Agriculture](https://www.analyticssteps.com/blogs/5-types-of-approaches-and-technologies-improve-agriculture-analytics) lacks organizational attention and support from bank loans and other welfare schemes.  A business analyst can also process the bank loan and farmers' welfare scheme for increasing agricultural inputs. It can give opportunities to farmers for individual income and growth. Another advancement is to give insight into irrigation, sowing,  harvest, and an area of land for the crop.

**2. Stock Marketing**

Business analyst improves the performance of the organization in terms of business process and profit by analyzing the variance in the market and updating the changing price or fluctuation in stock trends.

After analyzing the changes taking place in the stock market, he can provide the price list of shares of an item, relevant information regarding shareholder’s tracing of audience related to this market, etc.

The business analyst also makes flexible strategies and plans for future investment and earnings. For example, stocks are continuous trends in the market and change very frequently, so he works on the continuous trend dataset to implement strategies.

Business Analytics is surely a vast field to cover, many business analysts are now using [machine learning algorithms](https://www.analyticssteps.com/blogs/top-10-machine-learning-algorithms) and [natural language processing techniques](https://www.analyticssteps.com/blogs/top-10-natural-language-processing-nlp-trends-2021) in order to predict the growth or shrinking of stocks. Although the accuracy depends upon the quality of data being trained.

**3. Manufacturing Industry**

Business analysts seek to tackle the best possible items for manufacturing and supplying in the market. They investigate the detailed dataset to analyze and implement in business development.

For example, they describe the status to the supplier and product management team regarding the most manufactured product.

The role of technology like [IoT in the manufacturing industry](https://www.analyticssteps.com/blogs/role-iot-manufacturing-industry) shows the importance of technology and its uses.

A business analyst making use of such technologies can tell the highest number of customers for a specific product or service, the product’s performance, its demand elasticity (quality and quantity-wise), and product advertisement.

The right business analytics practice would do wonders for the manufacturing industry. For example, imagine how good it can be if you can predict when the sales would be high based on the factors such as location, season, and price! Surely, a company would know when to announce a discount offer on the basis of historical data and futuristic predictions.

**4. Finance Marketing**

Business Analysts utilize various analytical techniques and approaches to improve financially relevant issues such as fraud detection, risk mitigation, product pricing, marketing campaign optimization, financial planning, forecasting, etc.

These issues can be controlled by a business analyst. For example, a business analyst can find out the number of customers who are not making payments on time. Similarly, loan defaulters can be traced through a graphical representation that shows the defaulter’s age, gender, name, customer ID, etc.

**5. Medical Methodology**

In the medical or[healthcare department](https://www.analyticssteps.com/blogs/understanding-bioinformatics-application-machine-learning), the Business analyst makes predictions about the stock of medicine available in the hospital or medical store, the shipment of medicines in the local market, predictions related to disease, impacts of different medicines on same diseases, appointment and availability of doctor, arranging slots for patients, to a medicine available for cure.

For example, allotment of free slots to the patient considering the doctor’s working hours, duties of working staff in the hospital, etc.

Production of medicines can also be optimized by the business analyst. He proposes the strategies regarding production costs of medicines, areas of production and stock available, low cost, and high yield preparation methods.

**6. Customer Relation Management**

For every individual business, customer relationship management is the most important factor.

Business Analyst takes the required steps for a strong and healthy customer relationship with the organization. He helps to assemble emotional links with customers.

The business analyst helps in increasing productivity according to customer demand and variation in products with consumptions. He utilizes data to maintain the involvement of end to end-user and improves internal and external factors for better customer service and experience.

**7. Bond Marketing**

The bond market is not much behind the stock market in terms of sellers and buyers using online interfaces to buy and sell bonds. It gives people all the data they want and need to create data-driven trading strategies.

**8. Human Resources**

The application of business analytics in human resources can be understood by understanding its role in HR analytics.

The analysis of employees’ behavior has been recognized as [HR analytics](https://www.analyticssteps.com/blogs/what-hr-analytics-role-challenges-and-applications). Ultimately, business analysis plays a significant role in HR analytics as it provides crucial insights into the performance of a project.

The insights generated can then be utilized by HR to decide on the performance of employees on the project.

## Explain about model construction (14 marks)

## Building and Applying logistic regression models

**MODEL SELECTION**

## Competing goals:

* Should be complex enough to fit the data well.
* Should be simple to interpret – should smooth the data rather than overfitting it.

**Issue**: How to select a parsimonious (simple) model that fits the data well?

* Unrealistic to hope to find the *true* model for a real dataset.
* Part science, part statistics, part experience and part common sense.
* Less number of parameters leads to more precise estimates.
* Watch out for *collinearity* – correlation in the estimated coefficients. If two covariates are highly correlated, do not need both of them in the model.

## Indications of collinearity:

* Large standard errors.
* Look at the correlation matrix of the estimated coefficients. In R, use cor2cov(vcov(fit)),

where fit contains the glm fit.

## Indications of numerical instability:

* Error messages from the fitting program.
* Collinearity.
* Large standard errors.
* Zero or near-zero cell counts.
* Complete or near-complete separation. Complete separation means all zero responses appear at one combination of covariates and all one responses appear at another combination. No overlap in the covariates for the two responses. MLE does not exist in this case.

## Models building strategy:

**Step 1**: Use univariate analysis to identify important covariates – the ones that are at least moderately associated with response.

* One covariate at a time.
* Analyze contingency tables for each categorical covariate. Pay particular attention to cells with low counts. May need to collapse categories in a sensible fashion.
* Use nonparametric smoothing for each continuous covariate. Can also categorize the covariate and look at the plot of mean response (estimate of π) in each group against the group mid-point. To get a plot on logit scale, plot the logit transformation of this mean response. This plot also suggests the appropriate scale of the variable.
* an also fit logistic regression models with one covariate at a time and analyze the fits. In particular, look at the estimated coefficients, their standard errors and the likelihood ratio test for the significance of the coefficient.
* Rule of thumb: select all the variables whose p-value < 0.25 along with the variables of known clinical importance.

**Step 2**: Fit a multiple logistic regression model using the variables selected in step 1*.*

* Verify the importance of each variable in this multiple model using Wald statistic.
* Compare the coefficients of the each variable with the coefficient from the model containing only that variable.
* Eliminate any variable that doesn’t appear to be important, and fit a new model. Check if the new model is significantly different from the old model. If it is, then the deleted variable was important.
* Repeat this process of deleting, refitting and verifying until it appears that all the important variables are included in the model.
* At this point, add the variables into the model that were not selected in the original multiple model. Assess the joint significance of the variables that were not selected. This step is important as it helps to identify the confounding variables. Make changes in the model, if necessary.

At the end, we have the *preliminary main effects model* – it contains the important variables.

**Step 3**: Check the assumption of linearity in logit for each continuous covariate.

* Look at the smoothed plot of logit in step 1 against the covariate.
* If not linear, find a suitable transformation of the covariate so that the logit is roughly linear in the new variable.
* Try simple transformations such as power, log, etc. Also read about the method of fractional polynomials in the handout.

At the end, we have the *main effects model*.

**Step 4**: Check for interactions.

* Create a list of possible pairs of variables in the main effects model that have some scientific basis to interact with each other. This list may or may not consist of all possible pairs.
* Add the interaction term, one at a time, in the model containing all the main effects and assess its significance using the likelihood ratio test.
* Identify the significant interaction terms.

**Step 5**: Add the interactions found significant in step 4 to the main effects model and evaluate its fit.

* Look at the Wald tests and LR tests for the interaction terms.
* Drop any non-significant interaction.

At the end, we get our *preliminary final model*. We should now assess the overall goodness-of-fit of this model and perform model diagnostics.

**Example**: Read the analysis of UMARU impact study in the handout. You are expected to do the analysis for your project on similar lines.

**Another strategy**: Automatic stepwise selection procedure.

* Start with a list of important covariates obtained as before using the univariate analysis.
* *Forward selection*: Start with a simple model and add terms sequentially until further additions do not significantly improve the fit.
* *Backward elimination*: Start with a complex model and remove terms sequentially until a further deletion leads to a significantly poorer fit. (Generally preferred over forward selection).
* Other variants.
* Cannot trust the results.
* Can also use a penalized measure of model fit such as Akaike Information Criterion (AIC) instead of p-values. AIC = – 2(maximized log likelihood – # parameters in the model). Lower is better.

**Example**: Read section 6.1.3 for an example and Laura Thompson’s manual for R code.

## DETECTING LACK OF FIT

We have a model that we are reasonably satisfied with. The model fits well if the observed and fitted responses are close. With categorical covariates, it is likely that their # of distinct settings is (much) less than N. In other words, several subjects may have the same covariate setting.

**Goal**: Identify covariate patterns with lack of fit.

* J = # distinct covariate settings (patterns).
* mj = # of subjects in the j-th pattern. Then m1 + m2 + … + mJ = .
* oj = # of observed successes in the j-th covariate pattern.
* ej = fitted # of successes in the j-th covariate pattern = .
* Plot the observed versus fitted counts. If model fits well, the points should be close to the 45o line through origin*.* This method is effective only when J n
* With continuous covariates, it is likely that J n.

**Example**: Suppose there are 25 subjects in a study with 3 covariates – SEX, RACE and WEIGHT. We have 12 Males, 13 Females, 10 Whites, 8 Blacks and 7 Hispanics. Further, no two have the same weight.

1. Model has only SEX. Then J = , m1 = , m2 = .
2. Model has both SEX and RACE. Then J = .
3. Model has all three covariates. Then J = .

* Construct the following 2 x J contingency table and analyze the Pearson and deviance residuals for the observed and the fitted counts in this table.

Covariate Pattern

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | … | J | Total |
| Observed # y  1  Fitted # y  1 | *o*1  *e*1  *m*1ˆ1 | *o*2  *e*2  *m*2ˆ2 |  | *oJ*  *eJ*  *mJ* ˆ*J* |  |
| Observed # y  0  Fitted # y  0 | *m*1  *o*1 *m*1  *e*1 | *m*2  *o*2 *m*2  *e*2 |  | *mJ*  *oJ*  *mJ*  *eJ* |  |
| Total | *m*1 | *m*2 |  | *m*J | *N* |

1. Pearson residual:

*rj* 

*o j*  *mj* ˆ *j*

  *o j*   *mj*  *o j* 

*mj*ˆ *j* (1  ˆ *j* )

1. Deviance residual: *d j*

 2*o j* ln   (*m j*  *o j* ) ln  , where the sign +/- is same as

 *j*   *mj*  *e j* 

*e*



the sign of (*o j*  *e j* ) .

#   

 

* Both residuals are close to zero the observed and fitted counts are close.
* **Recall**: Roughly speaking, when *mj*

is large and the fitted model is correct, rj and dj are

approximately normal with mean zero and variance less than one. (Their standardized versions have variance one). Their absolute values larger than 2 or 3 indicate a possible lack of fit.

* Rule of thumb for normal approximation: Almost every cell should have a fitted count at least 5.

## INFLUENCE DIAGNOSTICS

***Goal***: Identify observations that have too much influence on fitted model.

* Delete one observation at a time, and look at the change in fit of the model and the estimates.
* Observation could be a single individual or a covariate pattern.
* Case-deletion diagnostics.

## Popular measures:

* Dfbetas = change in the estimated coefficients divided by its SE.
* Dffits = change in the fitted value divided by its SE.
* Cook’s distance = standardized sum of squares of change in all fitted values.
* covratio = change in covariance matrix of the estimates.
* Change in the Pearson or the LR 2 statistic.
* Most of these can be obtained in R using influence.measures(fit), where fit contains the glm fit.

## Using the diagnostics:

* Plot them against the estimated probabilities.
* Look for outlying points.

## What to do?

* Do not expect to identify many poorly fit or influential points when the model seems to fit well on overall goodness-of-fit measures (e.g., Hosmer-Lemeshow test).
* When there many such observations, one or more of following happened:
  + the logistic model is not a good approximation to the true relationship between  (*x*) and *x* .
  + an important covariate is missing
  + at least one of the covariates doesn’t have its correct scale in the model.
* Sometimes these problems can be alleviated by going back to the model building step.

Businesses are solving their liquidity crisis by enabling people to access more information than they used to have before. This helps them make their own reports and ideally make choices on the suitable bond for them.

Business analysts have started to gain almost the same amount of preference and attention given to bond traders.

.

Model Building (5 marks)

We've talked before about the "art" of model building. Unsurprisingly, there are many approaches to model building, but here is one strategy—consisting of seven steps—that is commonly used when building a regression model.

**The first step**

Decide on the type of model that is needed in order to achieve the goals of the study. In general, there are five reasons one might want to build a regression model. They are:

* For **predictive** reasons — that is, the model will be used to predict the response variable from a chosen set of predictors.
* For **theoretical** reasons — that is, the researcher wants to estimate a model based on a known theoretical relationship between the response and predictors.
* For **control** purposes — that is, the model will be used to control a response variable by manipulating the values of the predictor variables.
* For **inferential** reasons — that is, the model will be used to explore the strength of the relationships between the response and the predictors.
* For **data summary** reasons — that is, the model will be used merely as a way to summarize a large set of data by a single equation.

**The second step**

Decide which predictor variables and response variable on which to collect the data. Collect the data.

**The third step**

Explore the data. That is:

* On a univariate basis, check for outliers, gross data errors, and missing values.
* Study bivariate relationships to reveal other outliers, to suggest possible transformations, and to identify possible multicollinearities.

**The fourth step**

Randomly divide the data into a training set and a validation set:

* The **training set**, with at least 15-20 error degrees of freedom, is used to estimate the model.
* The **validation set** is used for cross-validation of the fitted model.

**The fifth step**

Using the training set, identify several candidate models:

* Use best subsets regression.
* Use stepwise regression, which of course only yields one model unless different alpha-to-remove and alpha-to-enter values are specified.

**The sixth step**

Select and evaluate a few "good" models:

* Select the models based on the criteria we learned, as well as the number and nature of the predictors.
* Evaluate the selected models for violation of the model conditions.
* If none of the models provide a satisfactory fit, try something else, such as collecting more data, identifying different predictors, or formulating a different type of model.

**The seventh and final step**

Select the final model:

* Compare the competing models by cross-validating them against the validation data.
* The model with a smaller mean square prediction error (or larger cross-validation *R*2) is a better predictive model.
* Consider residual plots, outliers, parsimony, relevance, and ease of measurement of predictors.

And, most of all, don't forget that there is not necessarily only **one** good model for a given set of data. There might be a few equally satisfactory models.

Blue Property Assumptions (14 marks)

**7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression**

[Ordinary Least Squares](https://statisticsbyjim.com/glossary/ordinary-least-squares/) (OLS) is the most common estimation method for linear models—and that’s true for a good reason. As long as your model satisfies the OLS assumptions for linear [regression](https://statisticsbyjim.com/glossary/regression-analysis/), you can rest easy knowing that you’re getting the best possible [estimates](https://statisticsbyjim.com/glossary/estimator/).

Regression is a powerful analysis that can analyze multiple variables simultaneously to answer complex research questions. However, if you don’t satisfy the OLS assumptions, you might not be able to trust the results.

What Does OLS Estimate and What are Good Estimates?

Regression analysis is like other [inferential methodologies](https://statisticsbyjim.com/basics/descriptive-inferential-statistics/). Our goal is to draw a [random sample](https://statisticsbyjim.com/glossary/sample/) from a [population](https://statisticsbyjim.com/glossary/population/) and use it to estimate the properties of that population.

In [regression analysis](https://statisticsbyjim.com/glossary/regression-analysis/), the [coefficients](https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/) in the regression equation are estimates of the actual population [parameters](https://statisticsbyjim.com/glossary/parameter/). We want these [coefficient](https://statisticsbyjim.com/glossary/regression-coefficient/) estimates to be the best possible estimates!

Suppose you request an estimate—say for the cost of a service that you are considering. How would you define a reasonable estimate?

1. The estimates should tend to be right on target. They should not be systematically too high or too low. In other words, they should be unbiased or correct on average.
2. Recognizing that estimates are almost never exactly correct, you want to minimize the discrepancy between the estimated value and actual value. Large differences are bad!

These two properties are exactly what we need for our coefficient estimates!

When your linear regression model satisfies the OLS assumptions, the procedure generates unbiased coefficient estimates that tend to be relatively close to the true population values (minimum variance). In fact, the Gauss-Markov theorem states that OLS produces estimates that are better than estimates from all other linear model estimation methods when the assumptions hold true.

The Seven Classical OLS Assumptions

Like many statistical analyses, ordinary least squares (OLS) regression has underlying assumptions. When these classical assumptions for linear regression are true, ordinary least squares produces the best estimates. However, if some of these assumptions are not true, you might need to employ remedial measures or use other estimation methods to improve the results.

Many of these assumptions describe properties of the error term. Unfortunately, the error term is a population value that we’ll never know. Instead, we’ll use the next best thing that is available—the [residuals](https://statisticsbyjim.com/glossary/residuals/). Residuals are the sample estimate of the error for each observation.

Residuals = Observed value – the [fitted value](https://statisticsbyjim.com/glossary/fitted-values/)

When it comes to checking OLS assumptions, assessing the residuals is crucial!

There are seven classical OLS assumptions for linear regression. The first six are mandatory to produce the best estimates. While the quality of the estimates does not depend on the seventh assumption, analysts often evaluate it for other important reasons that I’ll cover.

OLS Assumption 1: The regression model is linear in the coefficients and the error term

This assumption addresses the functional form of the model. In [statistics](https://statisticsbyjim.com/glossary/statistics/), a regression model is linear when all terms in the model are either the constant or a parameter multiplied by an independent variable. You build the model equation only by adding the terms together. These rules constrain the model to one type:

Y =\beta _{0} + \beta _{1}X_{1} + \beta _{2}X_{2} + \cdots + \beta _{k}X_{k} + \epsilon

In the equation, the betas (βs) are the parameters that OLS estimates. Epsilon (ε) is the random error.

In fact, the defining characteristic of linear regression is this functional form of the *parameters* rather than the ability to model curvature. Linear models can model curvature by including nonlinear *variables* such as polynomials and transforming exponential functions.

To satisfy this assumption, the correctly specified model must fit the linear pattern.

OLS Assumption 2: The error term has a population mean of zero

The error term accounts for the variation in the dependent variable that the independent variables do not explain. Random chance should determine the values of the error term. For your model to be unbiased, the average value of the error term must equal zero.

Suppose the average error is +7. This non-zero average error indicates that our model systematically underpredicts the observed values. Statisticians refer to systematic error like this as bias, and it signifies that our model is inadequate because it is not correct on average.

Stated another way, we want the expected value of the error to equal zero. If the expected value is +7 rather than zero, part of the error term is predictable, and we should add that information to the regression model itself. We want only random error left for the error term.

You don’t need to worry about this assumption when you include the constant in your regression model because it forces the mean of the residuals to equal zero. For more information about this assumption, read my post about the [regression constant](https://statisticsbyjim.com/regression/interpret-constant-y-intercept-regression/).

OLS Assumption 3: All independent variables are uncorrelated with the error term

If an independent variable is correlated with the error term, we can use the independent variable to predict the error term, which violates the notion that the error term represents unpredictable random error. We need to find a way to incorporate that information into the regression model itself.

This assumption is also referred to as exogeneity. When this type of [correlation](https://statisticsbyjim.com/glossary/correlation/) exists, there is endogeneity. Violations of this assumption can occur because there is simultaneity between the independent and dependent variables, omitted variable bias, or measurement error in the independent variables.

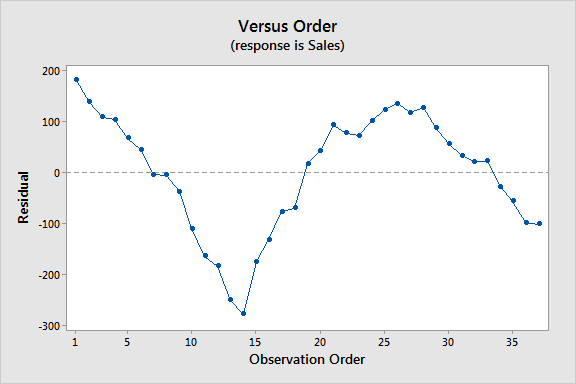
Violating this assumption biases the coefficient estimate. To understand why this bias occurs, keep in mind that the error term always explains some of the variability in the dependent variable. However, when an independent variable correlates with the error term, OLS incorrectly attributes some of the variance that the error term actually explains to the independent variable instead. For more information about violating this assumption, read my post about [confounding variables and omitted variable bias](https://statisticsbyjim.com/regression/confounding-variables-bias/).

OLS Assumption 4: Observations of the error term are uncorrelated with each other

One observation of the error term should not predict the next observation. For instance, if the error for one observation is positive and that systematically increases the probability that the following error is positive, that is a positive correlation. If the subsequent error is more likely to have the opposite sign, that is a negative correlation. This problem is known both as serial correlation and autocorrelation. Serial correlation is most likely to occur in time series models.

For example, if sales are unexpectedly high on one day, then they are likely to be higher than average on the next day. This type of correlation isn’t an unreasonable expectation for some subject areas, such as inflation rates, GDP, unemployment, and so on.

Assess this assumption by graphing the residuals in the order that the data were collected. You want to see randomness in the plot. In the graph for a sales model, there is a cyclical pattern with a positive correlation.



As I’ve explained, if you have information that allows you to predict the error term for an observation, you must incorporate that information into the model itself. To resolve this issue, you might need to add an independent variable to the model that captures this information. Analysts commonly use distributed lag models, which use both current values of the dependent variable and past values of independent variables.

For the sales model above, we need to add variables that explains the cyclical pattern.

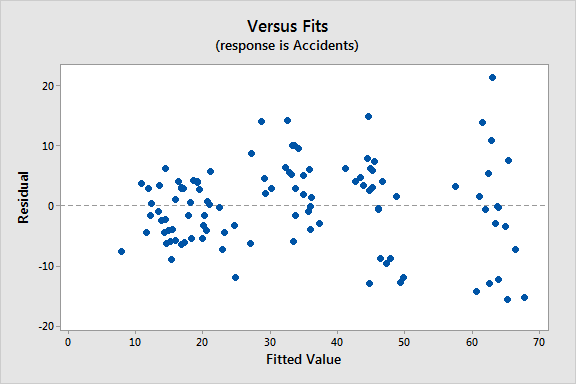
Serial correlation reduces the precision of OLS estimates. Analysts can also use time series analysis for time dependent effects.

An alternative method for identifying autocorrelation in the residuals is to[assess the autocorrelation function](https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/), which is a standard tool in time series analysis.

OLS Assumption 5: The error term has a constant variance (no heteroscedasticity)

The variance of the errors should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. This preferred condition is known as homoscedasticity (same scatter). If the variance changes, we refer to that as heteroscedasticity (different scatter).

The easiest way to check this assumption is to create a residuals versus fitted value plot. On this type of graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction. In the graph below, the spread of the residuals increases as the fitted value increases.



Heteroscedasticity reduces the precision of the estimates in OLS linear regression.

Note: When assumption 4 (no autocorrelation) and 5 (homoscedasticity) are both true, statisticians say that the error term is [independent and identically distributed (IID)](https://statisticsbyjim.com/basics/independent-identically-distributed-data/) and refer to them as spherical errors.

OLS Assumption 6: No independent variable is a perfect linear function of other explanatory variables

Perfect correlation occurs when two variables have a [Pearson’s correlation coefficient](https://statisticsbyjim.com/basics/correlations/) of +1 or -1. When one of the variables changes, the other variable also changes by a completely fixed proportion. The two variables move in unison.

Perfect correlation suggests that two variables are different forms of the same variable. For example, games won and games lost have a perfect negative correlation (-1). The temperature in Fahrenheit and Celsius have a perfect positive correlation (+1).

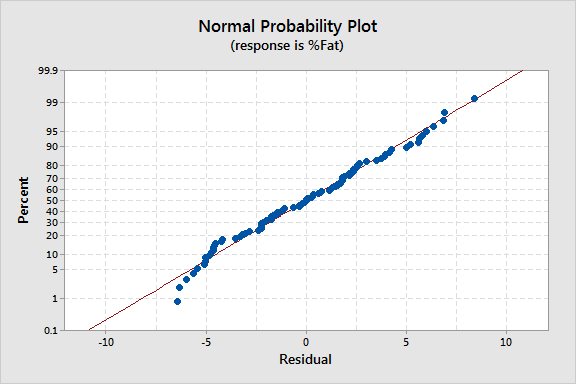
[Ordinary least squares](https://statisticsbyjim.com/glossary/ordinary-least-squares/) cannot distinguish one variable from the other when they are perfectly correlated. If you specify a model that contains independent variables with perfect correlation, your statistical software can’t fit the model, and it will display an error message. You must remove one of the variables from the model to proceed.

Perfect correlation is a show stopper. However, your statistical software can fit OLS regression models with imperfect but strong relationships between the independent variables. If these correlations are high enough, they can cause problems. Statisticians refer to this condition as multicollinearity, and it reduces the precision of the estimates in OLS linear regression.

OLS Assumption 7: The error term is normally distributed (optional)

OLS does not require that the error term follows a [normal distribution](https://statisticsbyjim.com/basics/normal-distribution/) to produce unbiased estimates with the minimum variance. However, satisfying this assumption allows you to perform statistical hypothesis testing and generate reliable confidence intervals and [prediction intervals](https://statisticsbyjim.com/glossary/prediction-intervals/).

The easiest way to determine whether the residuals follow a normal distribution is to assess a normal probability plot. If the residuals follow the straight line on this type of graph, they are normally distributed. They look good on the plot below!



If you need to obtain [p-values for the coefficient estimates](https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/) and the [overall test of significance](https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/), check this assumption!

Why You Should Care About the Classical OLS Assumptions

In a nutshell, your linear model should produce residuals that have a mean of zero, have a constant variance, and are not correlated with themselves or other variables.

If these assumptions hold true, the OLS procedure creates the best possible estimates. In statistics, estimators that produce unbiased estimates that have the smallest variance are referred to as being “efficient.” Efficiency is a statistical concept that compares the quality of the estimates calculated by different procedures while holding the sample size constant. OLS is the most efficient linear regression [estimator](https://statisticsbyjim.com/glossary/estimator/) when the assumptions hold true.

Another benefit of satisfying these assumptions is that as the sample size increases to infinity, the coefficient estimates converge on the actual population parameters.

If your error term also follows the normal distribution, you can safely use hypothesis testing to determine whether the independent variables and the entire model are statistically significant. You can also produce reliable confidence intervals and prediction intervals.

Knowing that you’re maximizing the value of your data by using the most efficient methodology to obtain the best possible estimates should set your mind at ease. It’s worthwhile checking these OLS assumptions! The best way to assess them is by using residual plots