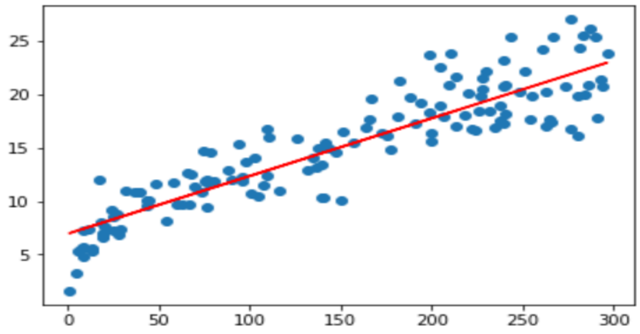
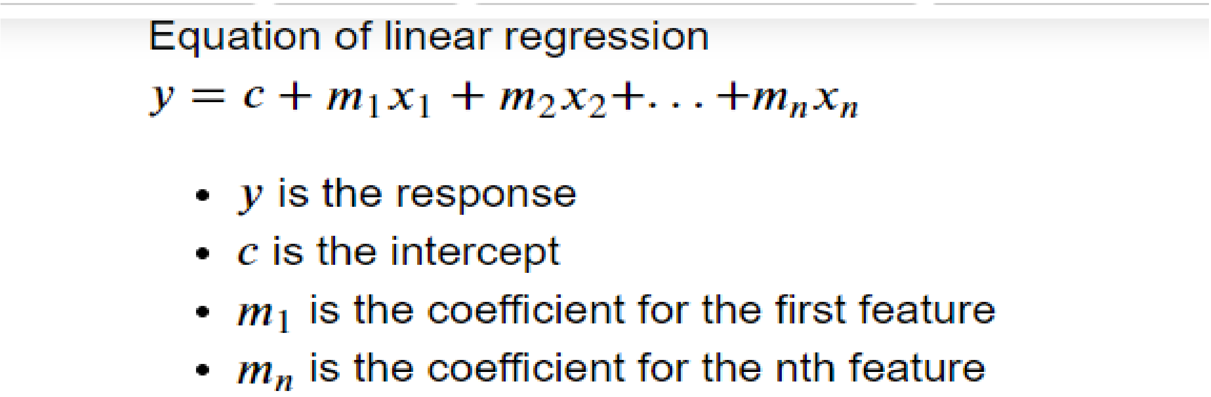
1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.



Equation of linear regression

$$y = c + m_1x_1 + m_2x_2 + \ldots + m_nx_n$$

- y is the response
- c is the intercept
- m_1 is the coefficient for the first feature
- m_n is the coefficient for the nth feature

While training the model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

How to update $_1$ and θ_2 values to get the best fit line?
Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the c and m values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

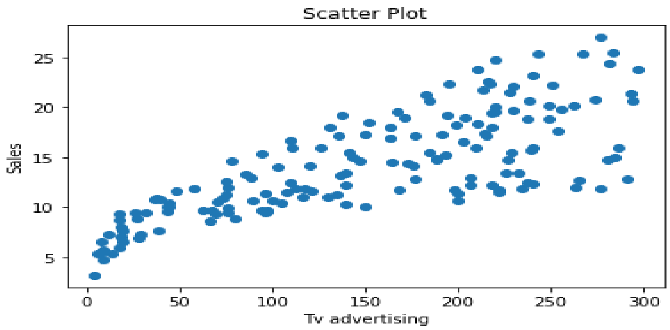
Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

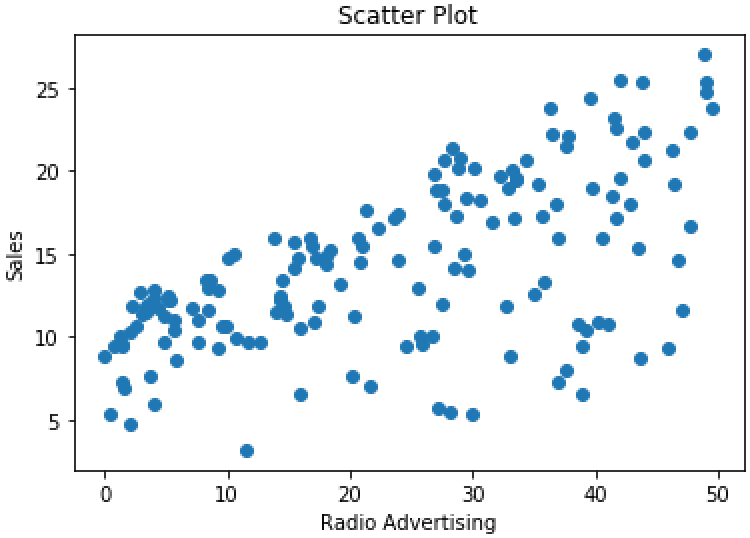
Gradient Descent:

To update c and m values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random c and m values and then iteratively updating the values, reaching minimum cost.

2. What are the assumptions of linear regression regarding residuals?

According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.

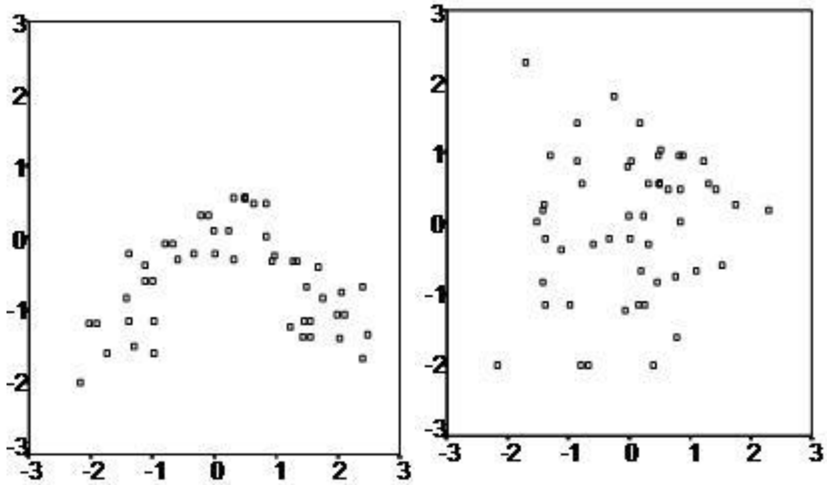




The first scatter plot of the feature TV vs Sales tells us that as the money invested on Tv advertisement increases the sales also increases linearly and the second scatter plot which is the feature Radio vs Sales also shows a partial linear relationship between them, although not completely linear.

Multiple Linear Regression Assumptions

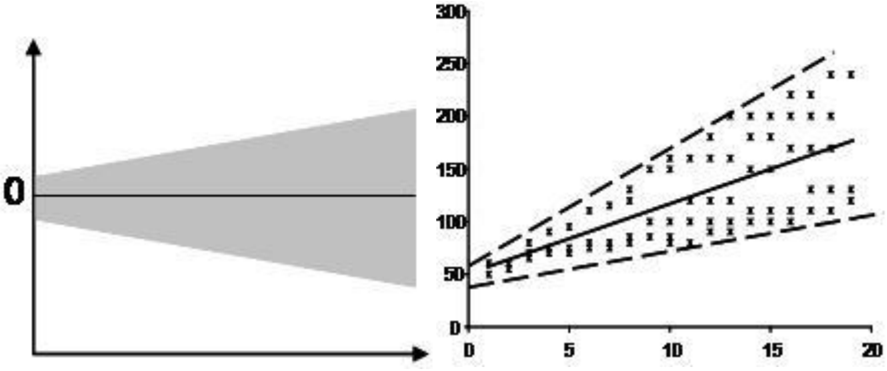
First, multiple linear regression requires the relationship between the independent and dependent variables to be linear. The linearity assumption can best be tested with scatterplots. The following two examples depict a curvilinear relationship (left) and a linear relationship (right).



Second, the multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. This assumption may be checked by looking at a histogram or a Q-Q-Plot. Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves.

Third, multiple linear regression assumes that there is no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be checked multiple ways:
1) Correlation matrix – When computing a matrix of Pearson's bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.
2) Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.
If multicollinearity is found in the data, one possible solution is to center the data. To center the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.
The last assumption of multiple linear regression is homoscedasticity. A scatterplot of residuals versus predicted values is good way to check for homoscedasticity. There should be no clear pattern in the distribution; if there is a cone-shaped pattern (as shown below), the data is heteroscedastic.

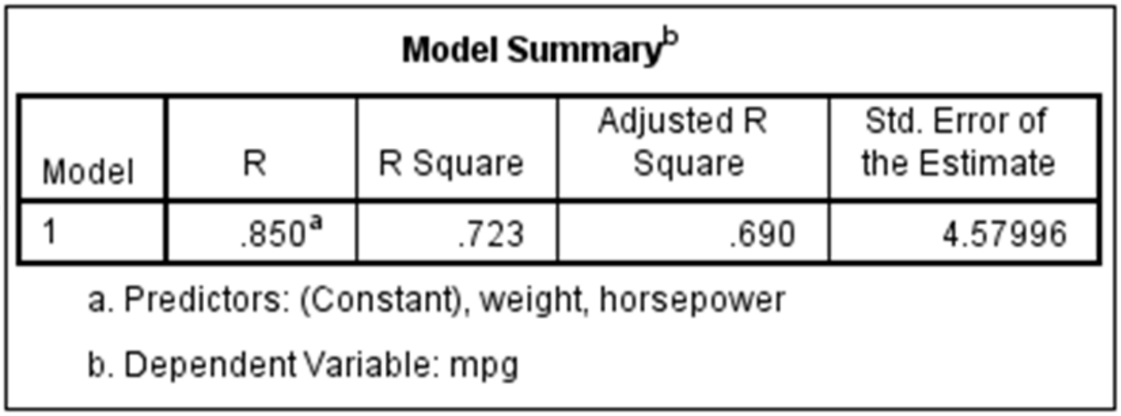


If the data are heteroscedastic, a non-linear data transformation or addition of a quadratic term might fix the problem.

3. What is the coefficient of correlation and the coefficient of determination?

Coefficient of correlation is "R" value which is given in the summary table in the Regression output. R square is also called coefficient of determination. Multiply R times R to get the R square value. In other words, Coefficient of Determination is the square of Coefficient of Correlation.
R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

It is easy to explain the R square in terms of regression. It is not so easy to explain the R in terms of regression.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.850[a]	.723	.690	4.57996

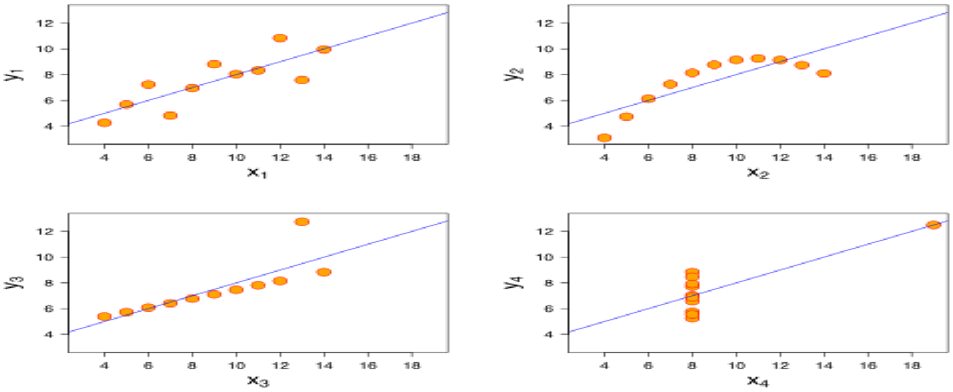
a. Predictors: (Constant), weight, horsepower

b. Dependent Variable: mpg

Coefficient of Correlation is the R value i.e. .850 (or 85%). Coefficient of Determination is the R square value i.e. .723 (or 72.3%). R square is simply square of R i.e. R times R.
Coefficient of Correlation: is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated, then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related). Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That’s why R square is a better term. You can explain R square for both simple linear regressions and for multiple linear regressions.

4. Explain the Anscombe's quartet in detail.

All four sets are identical when examined using simple summary statistics but vary considerably when graphed.



Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.[1]

For all four datasets:

Property	Value
Mean of x in each case:	9 (exact)
Variance of x in each case:	11 (exact)
Mean of y in each case:	7.50 (to 2 decimal places)
Variance of y in each case:	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case:	0.816 (to 3 decimal places)
Linear regression line in each case:	y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant. In the third graph (bottom left), the distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The x values are the same for the first three datasets

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

A procedure to generate similar data sets with identical statistics and dissimilar graphics has since been developed.

5. What is Pearson’s R?

Pearson’s correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Feature scaling (also known as **data normalization**) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

Scaling

Normalization rescales *(also called **min-max scaling**)*, transform the data such that the features are within a specific range e.g. [0, 1] .This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization rescales data to have a mean (μμ) of 0 and standard deviation (σσ) of 1 (unit variance).

$$X_{changed} = \frac{X - \mu}{\sigma}$$

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features. [0]
VIF = 1 / (1 – R^2)
When R^2 reaches 1, VIF reaches infinity
We try to remove features for which VIF > 5

8. What is the Gauss-Markov theorem?

A theorem that proves that if the error terms in a multiple regression have the same variance and are uncorrelated, then the estimators of the parameters in the model produced by least squares estimation are better (in the sense of having lower dispersion about the mean) than any other unbiased linear estimator.
This is pretty much considered the “big boy” reason least squares fitting can be considered a good implementation of linear regression.
Suppose you are building a model of the form:
y(i) = B . x(i) + e(i)
where B is a vector (to be inferred), i is an index that runs over the available data (say 1 through n), x(i) is a per-example vector of features, and y(i) is the scalar quantity to be modeled. Only x(i) and y(i) are observed. The e(i) term is the un-modeled component of y(i) and you typically hope that the e(i) can be thought of unknowable effects, individual variation, ignorable errors, residuals, or noise. How weak/strong assumptions you put on the e(i) (and other quantities) depends on what you know, what you are trying to do, and which theorems you need to meet the pre-conditions of. The Gauss-Markov theorem assures a good estimate of B under weak assumptions.

9. Explain the gradient descent algorithm in detail.

Most machine learning algorithms perform predictive modeling by minimizing an objective function, thereby learning the weights that must be applied to the testing data in order to obtain the predicted labels. The simplest objective (loss) function is the Sum of Squared Errors (SSE) function which we will denote as J(**w**):

$$J(\boldsymbol{w}) = \frac{1}{2} \sum_{(x,y)\in D} (y - prediction(x;\boldsymbol{w}))^2$$

Here, x represents the features, y the labels, D the training dataset containing features and labels, and **w** are the weights learned from the model by minimizing the objective function. The objective function is often minimized using the gradient descent (GD) algorithm. In the GD method, the weights are updated according to the following procedure:

$$\boldsymbol{w} = \boldsymbol{w} - \eta\, \nabla J(\boldsymbol{w})$$

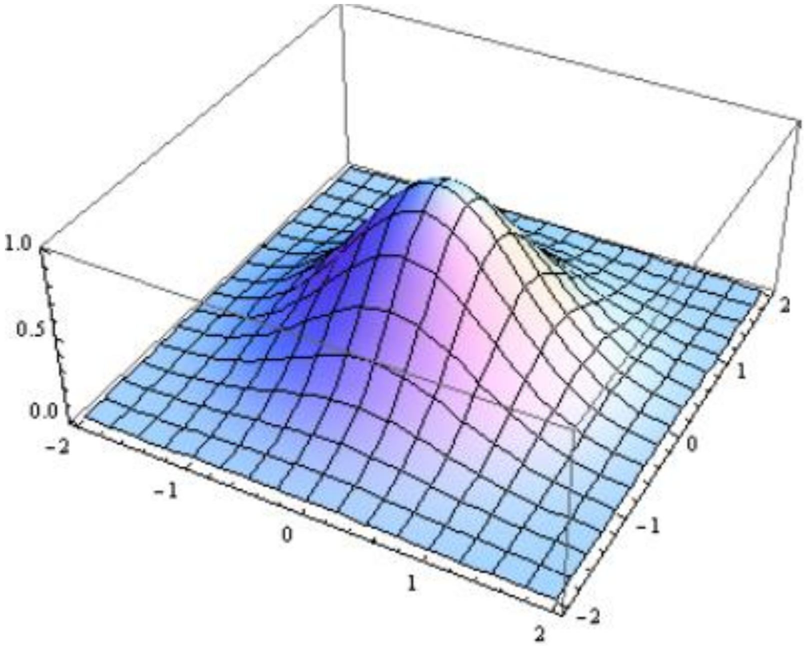
i.e., in the direction opposite to the gradient. Here, eta is a small positive constant referred to as the learning rate.

But why does the GD algorithm work?

Why the GD algorithm works

To illustrate, we consider a simple example, namely, the 2D Gaussian function. We perform calculations using Mathematica software.

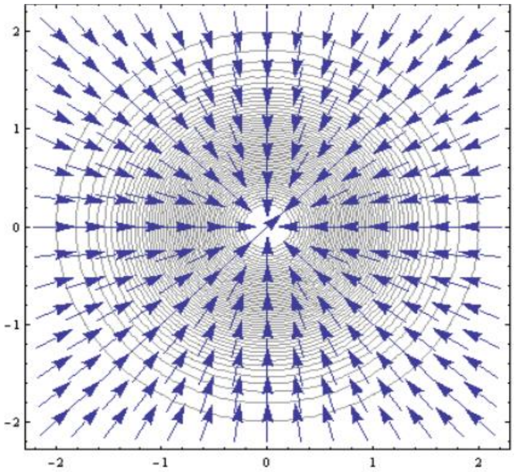
Plot3D[Exp[-(x² + y²)], {x, -2, 2}, {y, -2, 2}, PlotRange -> All]



3D plot of the Gaussian function. This function has a maximum at the origin, i.e., at (0, 0).

We see that the function has a maximum value at (0,0). Now let us generate a contour plot of the function and superimpose on it the unit vector in the direction of the gradient vector:

p1 = ContourPlot[Exp[-(x² + y²)], {x, -2, 2}, {y, -2, 2}];
p2 = VectorPlot[{-x/Sqrt[x² + y²] , -y/Sqrt[x² + y²]}, {x, -2,2}, {y, -2, 2}];Show[p1, p2]



Contour plot of the Gaussian function and vector fields of the gradient vector.

We observe that the vector fields of the unit vectors (arrows) of the gradient point towards the origin (0,0), where the function attains its maximum value. So we see that the gradient vector always points in the direction of the maximum of a function.

Hence the following rules apply:

To **Maximize** a function of several variables, we take steps in the direction of the gradient vector.

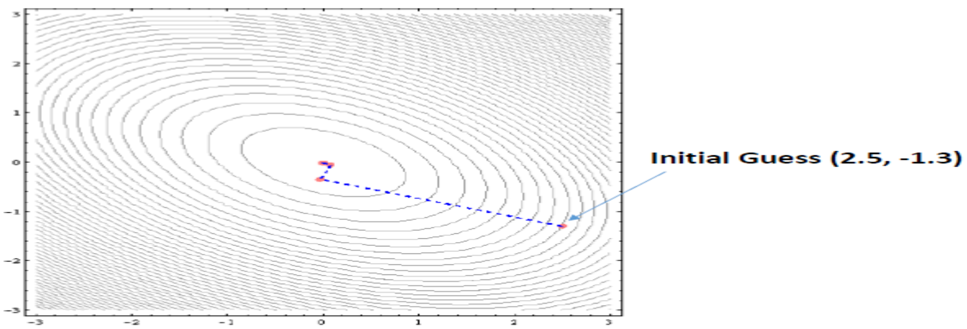
To **Minimize** a function of several variables, we take steps opposite the direction of the gradient vector.

Example of GD algorithm

Suppose we want to minimize the objective function:

$$J(w_x, w_y) = w_x^2 + w_x w_y + 1.5 w_y^2$$

Clearly, this function has a global minimum of 0 at the point (w_x =0, w_y = 0). Applying the GD algorithm with some initial guess (w_x = 2.5, w_y = -1.3), we can show with a few lines of code that the algorithm converges to the correct minimum, that is to the point (0,0), as shown here:

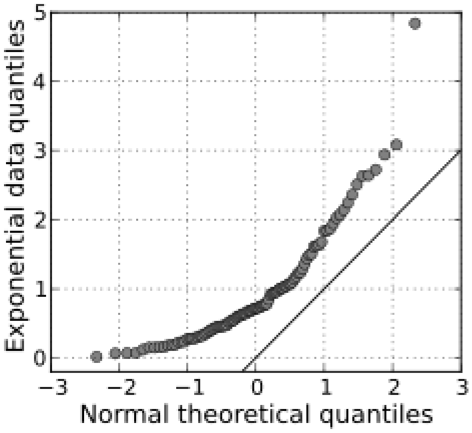


Initial Guess (2.5, -1.3)

Minimization of a simple objective function for illustrative purposes.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



A Q Q plot showing the 45-degree reference

The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a **normal quantile-quantile (QQ) plot.** The points are not clustered on the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

References: -

1. http://Learn.upgrad.com
2. https://www.geeksforgeeks.org/ml-linear-regression/
3. https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9
4. https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/#:~:targetText=Multivariate%20Normality%E2%80%93Multiple%20regression%20assumes,Inflation%20Factor%20(VIF)%20values.
5. http://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-is-the-difference-between-coefficient-of-determination-and-coefficient-of-correlation/#:~:targetText=Coefficient%20of%20correlation%20is%20%E2%80%9CR,table%20in%20the%20Regression%20output.&targetText=In%20other%20words%20Coefficient%20of,all%20the%20x%20variables%20together.

6. https://bobsleanlearning.wordpress.com/2013/01/02/anscombes-quartet/
7. https://www.statisticssolutions.com/pearsons-correlation-coefficient/
8. http://www.win-vector.com/blog/2014/08/reading-the-gauss-markov-theorem/
9. https://medium.com/towards-artificial-intelligence/machine-learning-how-the-gradient-descent-algorithm-works-61682d8570b6
10. https://www.statisticshowto.datasciencecentral.com/q-q-plots/#:~:targetText=The%20purpose%20of%20Q%20Q%20plots,the%2045%20degree%20reference%20line.
11. https://www.quora.com/What-is-gradient-descent-algorithm