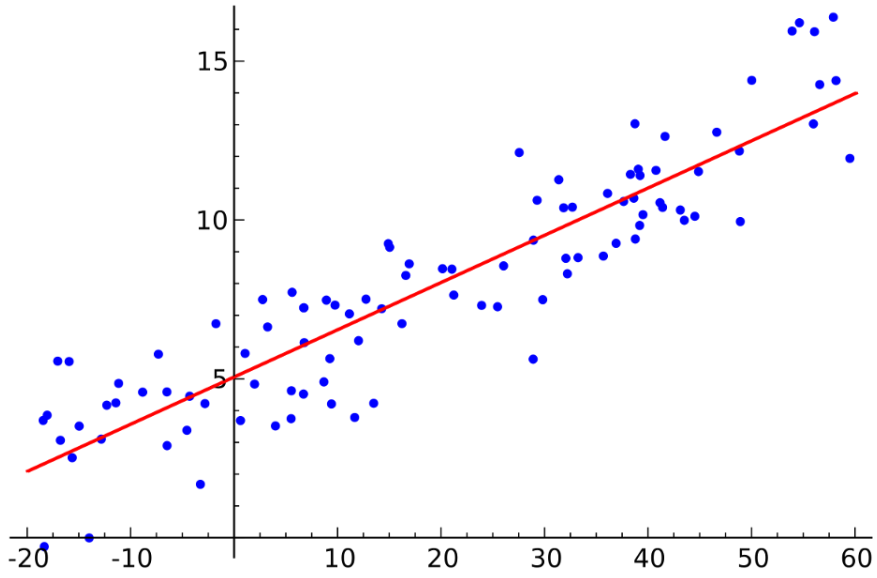
ASSIGNMENT – 13th Jan 2020

1. LINEAR REGRESSION ALGORITHM IN DETAIL:

- Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression.

1. Simple Linear
2. Multiple Linear Regression



SIMPLE LINEAR REGRESSION:

- Simple linear regression is useful for finding relationship between two continuous variables
- One or more predictors or independent variables and other is response or dependent variable
- Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other
- Before attempting to fit a linear model to observed data, we should decide whether there is a relationship between the variables of interest
- A scatterplot can be a helpful tool in determining the strength of the relationship between two variables.
- For example, we have a dataset which contains information about relationship between 'number of hours studied' and 'marks obtained'. Many students have been observed and their hours of study and grade are recorded. This will be our training data. Goal is to design a model that can predict marks if given the number of hours studied. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used for any new data. That is, if we give number of hours studied by a student as an input, our model should predict their mark with minimum error.

Y (pred) = b0 + b1*x

- The values b0 and b1 must be chosen so that they minimize the error. If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error.

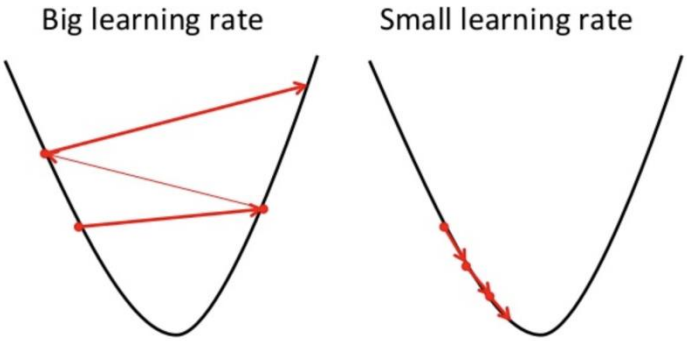
$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

- Cost Function –
  - The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. Since we want the best values for a_0 and a_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

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

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

- Gradient Descent –
  - Gradient descent is a method of updating a_0 and a_1 to reduce the cost function (MSE). The idea is that we start with some values for a_0 and a_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.



  - Your objective is to reach the bottom of the pit. There is a catch, you can only take a discrete number of steps to reach the bottom.
  - If you decide to take one step at a time you would eventually reach the bottom of the pit but this would take a longer time. If you choose to take longer steps each time, you would reach sooner but, there is a chance that you could overshoot the bottom of the pit and not exactly at the bottom

- For model with one predictor,

$$b_0 = \bar{y} - b_1 \bar{x}$$

Intercept Calculation

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Co-efficient Formula

- Exploring ‘b1’
  - If b1 > 0, then x (predictor) and y (target) have a positive relationship. That is increase in x will increase y.
  - If b1 < 0, then x (predictor) and y (target) have a negative relationship. That is increase in x will decrease y.
- Exploring ‘bo’
  - If the model does not include x=0, then the prediction will become meaningless with only b0.
  - If the model includes value 0, then ‘b0’ will be the average of all predicted values when x=0.
  - The value of b0 guarantee that residual have mean zero. If there is no ‘b0’ term, then regression will be forced to pass over the origin. Both the regression co-efficient and prediction will be biased.

- Residual Analysis
  - Randomness and unpredictability are the two main components of a regression model.

Prediction = Deterministic + Statistic

  - Deterministic part is covered by the predictor variable in the model. Stochastic part reveals the fact that the expected and observed value is unpredictable. The information using which we were not able to cover is residual information.
- R- squared value
  - This value ranges from 0 to 1. Value ‘1’ indicates predictor perfectly accounts for all the variation in Y. Value ‘0’ indicates that predictor ‘x’ accounts for no variation in ‘y’.

Conduct Linear Regression Model:

We use scikit learn to import the linear regression model. We fit the model on the training data and predict the values for the testing data. We use R2 score to measure the accuracy of our model. Let the example be 90.01.

We initialize the value 0.0 for b0 and b1. For 1000 epochs we calculate the cost, and using the cost we calculate the gradients, and using the gradients we update the values of b0 and b1. After

1000 epochs, we would've obtained the best values for b0 and b1 and hence, we can formulate the best fit straight line.

The test set contains 300 samples, therefore we have to reshape b0 and b1 from 700x1 to 300x1. Now, we can just use the equation to predict values in the test set and obtain the R2 score.

As R2 previous model was 90.01 now if it is 91.7, it indicates that the model is performing better than previous. It mean that 91.7% of data was explained using this particular model.

2. ASSUMPTIONS IN LINEAR REGRESSION REGARDING RESIDUALS:

1. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as **Autocorrelation**.
2. The error terms must have constant variance. This phenomenon is known as **homoscedasticity**. The presence of non-constant variance is referred to heteroscedasticity.
3. The error terms must be **normally distributed**.

1. Autocorrelation: The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

If this happens, it causes confidence intervals and prediction intervals to be narrower. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients. Let's understand narrow prediction intervals with an example:

For example, the least square coefficient of X^1 is 15.02 and its standard error is 2.08 (without autocorrelation). But in presence of autocorrelation, the standard error reduces to 1.20. As a result, the prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).

Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.

2. Heteroscedasticity: The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

How to check: If heteroscedasticity exists, the plot would exhibit a funnel shape pattern

3. Normal Distribution of error terms: If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

3. COEFFICIENT OF CORRELATION & COEFFICIENT OF DETERMINATION:

1. COEFFICIENT OF CORRELATION

- 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 coefficient of determination shows percentage variation in y which is explained by all the x variables together.
- Higher the value, better would be the explanation of variation
- The value of coefficient correlation would be ranging from 0 to 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.

2. COEFFICIENT OF DETERMINATION:

- Coefficient of determination is the degree of relationship between two variables 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 also for multiple linear regressions.

4. ANSCOMBE's QUARTET:

- The purpose of a scatter plot is to visually communicate the relationship between numerical (interval or ratio scale) variables.
- While a correlation coefficient is a statistic that can be used to describe the strength of a linear relationship, a visual can better describe the nature of relationship and the behavior of the underlying variables.
- Anscombe's quartet is a classic example of the drawback to just reporting correlation. Frank Anscombe illustrated in his 1973 that how a set of four different pairs of variables can deliver the same correlation coefficient, while the relationships between each pair are completely different.
- **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 analysing it and the effect of outliers on statistical properties.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Figure 1 Dataset values for each dataset consist of eleven points

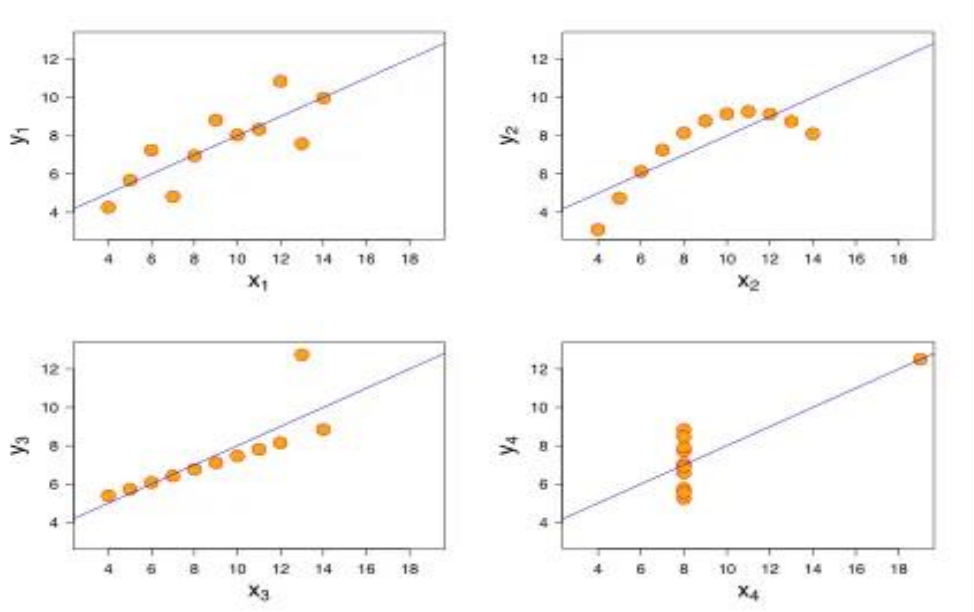


Figure 2 Plot showing different stories for similar dataset

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 analyse 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

5. PEARSON's CORRELATION COEFFICIENT:

- Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables
- A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related.
- 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.
- Assumptions:
  - Independent of case: Cases should be independent to each other.
  - Linear relationship: Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
  - Homoscedasticity: the residuals scatterplot should be roughly rectangular-shaped.
- Properties:

- Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.
- Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
- Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

- Degree of correlation:
  - Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
  - High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
  - Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
  - Low degree: When the value lies below + .29, then it is said to be a small correlation.
  - No correlation: When the value is zero.

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

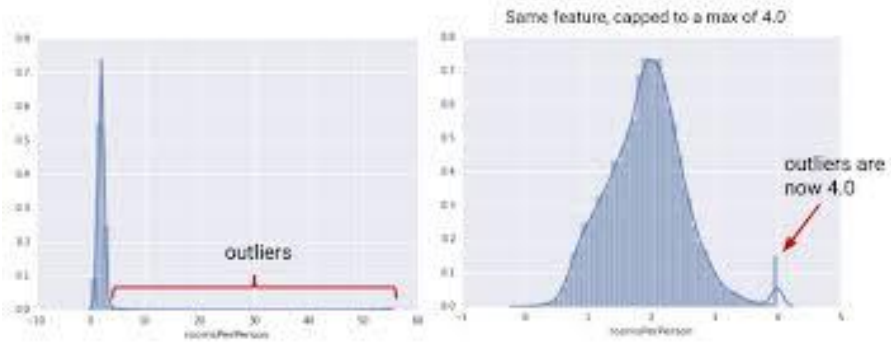
Definition: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Reason for scaling: Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem.
If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.
To supress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Difference between normalized scaling and standardized scaling:

Normalisation:

The point of normalization is to change your observations so that they can be described as a normal distribution.



The word "normalization" is used informally in statistics, and so the term *normalized data* can have multiple meanings. In most cases, when you normalize data you eliminate the units of measurement for data, enabling you to more easily compare data from different places. Some of the more common ways to normalize data include:

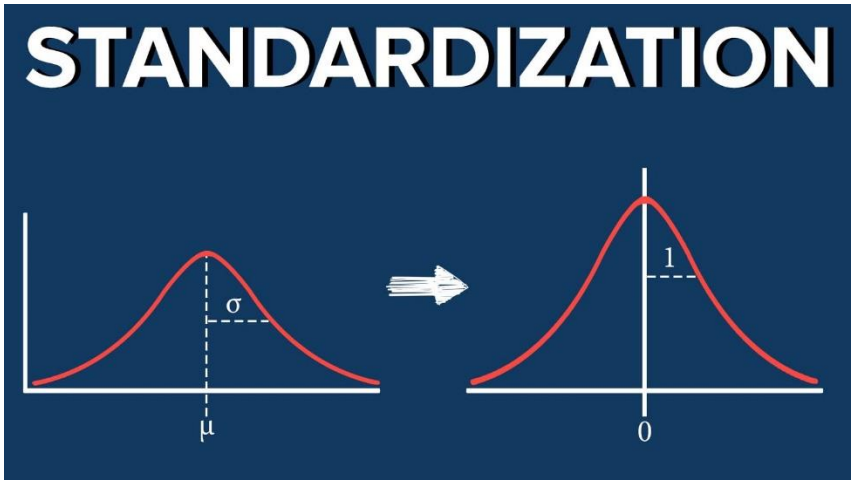
- Transforming data using a z-score or t-score. This is usually called *standardization*. In the vast majority of cases, if a statistics textbook is talking about normalizing data, then this is the definition of "normalization" they are probably using.
- Rescaling data to have values between 0 and 1. This is usually called *feature scaling*. One possible formula to achieve this is:

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

- Standardizing residuals: Ratios used in regression analysis can force residuals into the shape of a normal distribution.
- Normalizing Moments using the formula μ/σ.
- Normalizing vectors (in linear algebra) to a norm of one. Normalization in this sense means to transform a vector so that it has a length of one.

Standardisation:

The result of standardization (or Z-score normalization) is that the features will be rescaled so that they'll have the properties of a standard normal distribution with μ=0 and σ=1



Where μ is the mean (average) and σ is the standard deviation from the mean; standard scores (also called *z* scores) of the samples are calculated as follows:

- $z=x-\mu/\sigma$

Normalization vs. Standardization

The terms *normalization* and *standardization* are sometimes used interchangeably, but they usually refer to different things. *Normalization* usually means to scale a variable to have a values between 0 and 1, while *standardization* transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a z-score, and data points can be standardized with the following formula:

$$z_i = \frac{x_i - \bar{x}}{s}$$

A z-score standardizes variables.

Where:

- x_i is a data point (x_1, $x_2 \ldots x_n$).
- $\bar{x}$ is the sample mean.
- s is the sample standard deviation.

Z-scores are very common in statistics. They allow you to compare different sets of data and to find probabilities for sets of data using standardized tables (called z-tables).

7. VIF:

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.

How the VIF is computed

The *standard error* of an *estimate* in a *linear regression* is determined by four things:

- The overall amount of noise (error). The more noise in the data, the higher the standard error.
- The variance of the associated predictor variable. The greater the variance of a predictor, the smaller the standard error (this is a *scale* effect).
- The sampling mechanism used to obtain the data. For example, the smaller the sample size with a simple random sample, the bigger the standard error.
- The extent to which a predictor is correlated with the other predictors in a model.

The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the *R-squared* statistic of the regression where the predictor of interest is predicted by all the other predictor variables ( ). The *variance inflation* for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

Some statistical software use *tolerance* instead of VIF, where tolerance is:

$1 - R^2 = \frac{1}{VIF}$.

The VIF can be applied to any type of predictive model (e.g., CART, or deep learning). A generalized version of the VIF, called the *GVIF*, exists for testing sets of predictor variables and generalized linear models.

How to interpret the VIF

A VIF can be computed for each predictor in a predictive model. A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high. These numbers are just rules of thumb; in some contexts a VIF of 2 could be a great problem (e.g., if estimating *price elasticity*), whereas in straightforward predictive applications very high VIFs may be unproblematic.

If all the independent variables are orthogonal (independent) to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

8. Gauss Markov Theorem

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.

There are five Gauss Markov assumptions (also called *conditions*):

1. Linearity: the parameters we are estimating using the OLS method must be themselves linear.
2. Random: our data must have been randomly sampled from the population.
3. Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
4. Exogeneity: the regressors aren't correlated with the error term.
5. Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

Purpose of the Assumptions

The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

In practice, the Gauss Markov assumptions are rarely all met perfectly**,** but they are still useful as a benchmark, and because they show us what 'ideal' conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

The Gauss-Markov theorem famously states that OLS is BLUE. BLUE is an acronym for the following:

Best Linear Unbiased Estimator

In this context, the definition of "best" refers to the minimum variance or the narrowest sampling distribution. More specifically, when your model satisfies the assumptions, OLS coefficient estimates follow the tightest possible sampling distribution of unbiased estimates compared to other linear estimation methods.

What Does OLS Estimate?

Regression analysis is like any other inferential methodology. Our goal is to draw a random sample from a population and use it to estimate the properties of that population. In regression analysis, the coefficients in the equation are estimates of the actual population parameters.

The notation for the model of a population is the following:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

The betas (β) represent the population parameter for each term in the model. Epsilon (ε) represents the random error that the model doesn't explain. Unfortunately, we'll never know these population values because it is generally impossible to measure the entire population. Instead, we'll obtain estimates of them using our random sample.

The notation for an estimated model from a random sample is the following:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k + e$$

The hats over the betas indicate that these are parameter estimates while e represents the residuals, which are estimates of the random error.

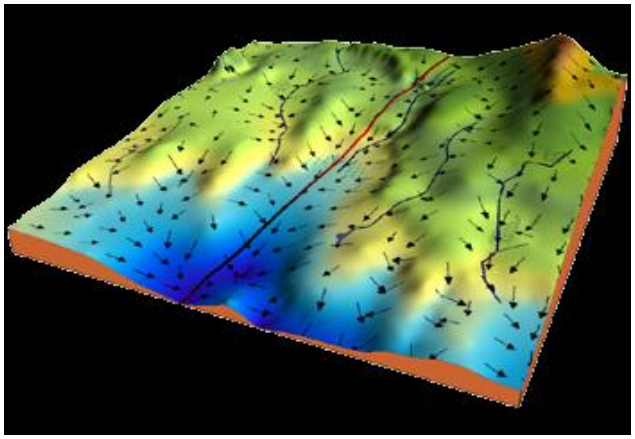
Typically, statisticians consider estimates to be useful when they are unbiased (correct on average) and precise (minimum variance).

9. GRADIENT DESCENT ALGORITHM:

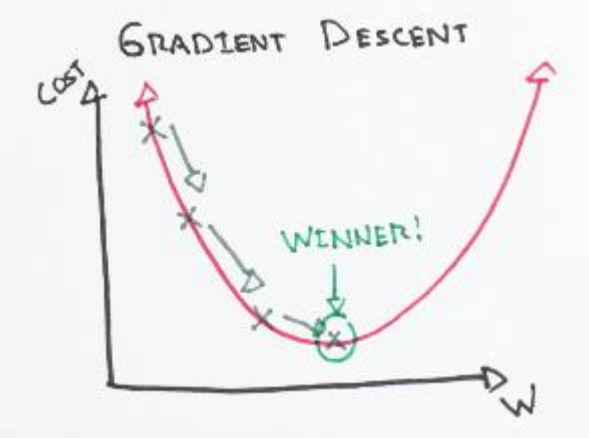
Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

Introduction

Consider the 3-dimensional graph below in the context of a cost function. Our goal is to move from the mountain in the top right corner (high cost) to the dark blue sea in the bottom left (low cost). The arrows represent the direction of steepest descent (negative gradient) from any given point–the direction that decreases the cost function as quickly as possible



Starting at the top of the mountain, we take our first step downhill in the direction specified by the negative gradient. Next we recalculate the negative gradient (passing in the coordinates of our new point) and take another step in the direction it specifies. We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum



Learning rate

The size of these steps is called the *learning rate*. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more

precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

Cost function

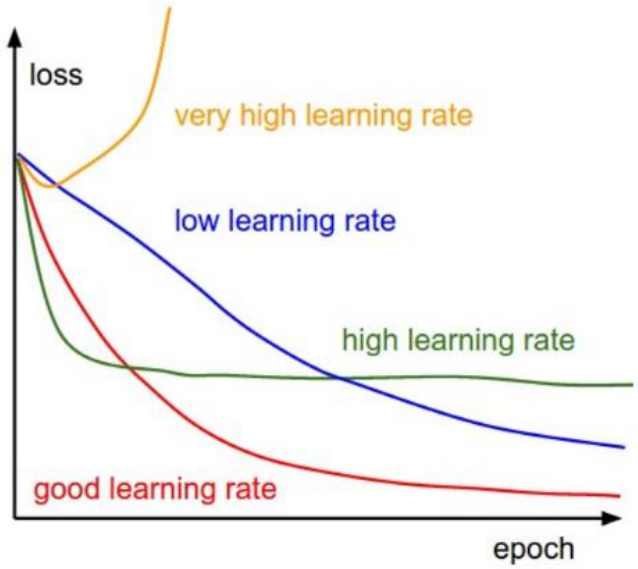
A Loss Functions tells us "how good" our model is at making predictions for a given set of parameters. The cost function has its own curve and its own gradients. The slope of this curve tells us how to update our parameters to make the model more accurate.

Step-by-step

Now let's run gradient descent using our new cost function. There are two parameters in our cost function we can control: w (weight) and b (bias). Since we need to consider the impact each one has on the final prediction, we need to use partial derivatives. We calculate the partial derivatives of the cost function with respect to each parameter and store the results in a gradient.

1. Initialize weight *w* and bias *b* to any random numbers.
2. Pick value for learning rate.
    - If learning rate is small, it would take time to converge.
    - If learning rate is large, it will fail to converge and overshoot.

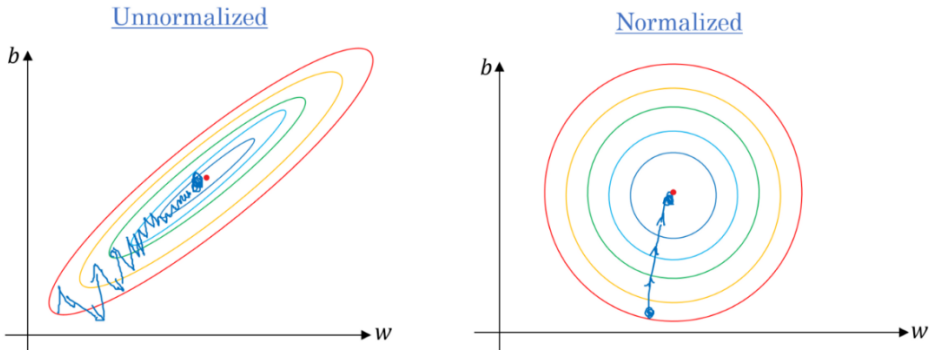
Therefore above procedure wholly depends on the deciding factor of learning rate. If we take less value for learning rate, we reach the converging point due to which the step seems costly. When we consider the large value for learning rate, it will fail to converge. Commonly used learning rate value is 0.001, 0.003,0.01,0.03,0.1,0.3



The above plot the cost function against different values of learning rate. Pick the right value of learning rate so that the learning algorithm converges.

3. Make sure to scale the data if it's on a very different scale. If we don't scale the data, the level curves would be narrower and taller which means it would take longer time to converge. The scaling of data have to done so that $\mu = 0$ and $\sigma = 1$. Below is the formula for scaling each example:

$$\frac{x_i - \mu}{\sigma} \qquad (1)$$



4. On each iteration, take the partial derivative of the cost function *J(w)* w.r.t each parameter (gradient):

$$\frac{\partial}{\partial w} J(w) = \nabla_w J \tag{2}$$

$$\frac{\partial}{\partial b} J(w) = \nabla_b J \tag{3}$$

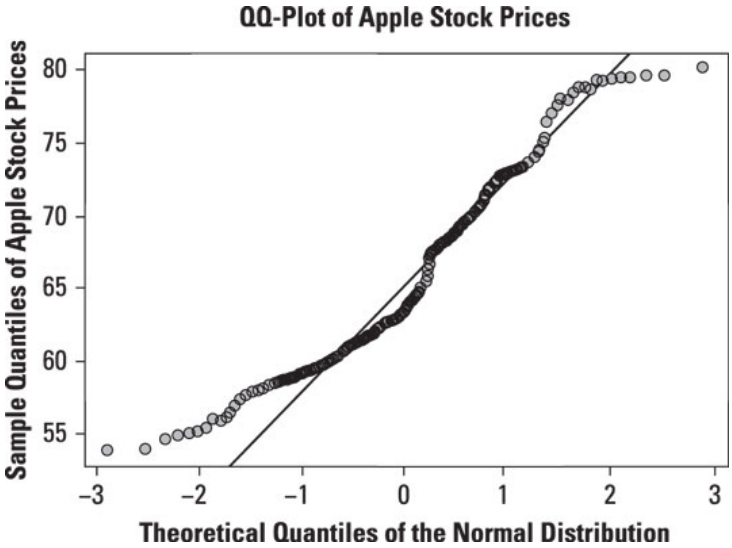
The update equations are:

$$w = w - \alpha \nabla_w J \tag{4}$$

$$b = b - \alpha \nabla_b J \tag{5}$$

10. Q-Q PLOT:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A *quantile-quantile plot* (also known as a *QQ-plot*) is another way you can determine whether a dataset matches a specified probability distribution. QQ-plots are often used to determine whether a dataset is *normally* distributed. Graphically, the QQ-plot is very different from a histogram. As the name suggests, the horizontal and vertical axes of a QQ-plot are used to show *quantiles.*
Quartiles divide a dataset into four equal groups, each consisting of 25 percent of the data. But there is nothing particularly special about the number four.
Another popular type of quantile is the *percentile*, which divides a dataset into 100 equal groups. For example, the 30th percentile is the boundary between the smallest 30 percent of the data and the largest 70 percent of the data. The median of a dataset is the 50th percentile of the dataset. The 25th percentile is the first quartile, and the 75th percentile the third quartile.
With a QQ-plot, the quantiles of the sample data are on the vertical axis, and the quantiles of a specified probability distribution are on the horizontal axis. The plot consists of a series of points that show the relationship between the actual data and the specified probability distribution. If the elements of a dataset perfectly match the specified probability distribution, the points on the graph will form a 45 degree line.
For example, this figure shows a normal QQ-plot for the price of Apple stock from January 1, 2013 to December 31, 2013.



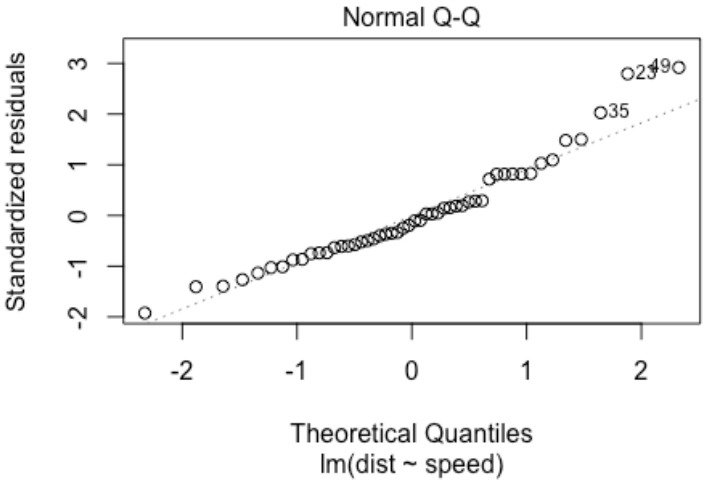
The QQ-plot shows that the prices of Apple stock do not conform very well to the normal distribution. In particular, the deviation between Apple stock prices and the normal distribution seems to be greatest in the lower left-hand corner of the graph, which corresponds to the *left tail* of the normal distribution. The discrepancy is also noticeable in the upper right-hand corner of the graph, which corresponds to the *right tail* of the normal distribution.

The graph shows that the smallest prices of Apple stock are not small enough to be consistent with the normal distribution; similarly, the largest prices of Apple stock are not large enough to be consistent with the normal distribution. This shows that the tails of the Apple stock price distribution are too "thin" or "skinny" compared with the normal distribution. The conclusion to be drawn from this is that the Apple stock prices are *not* normally distributed.

IMPORTANCE OF Q-Q PLOT:

QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, you can't assume your estimator is Gaussian either, so the standard confidence intervals and significance tests are invalid. i.e., if we have to compare the error term with quantiles of normal distribution If the quantiles of the error terms are near enough to the quantiles of the corresponding values computed from the normal distribution, then we might begin to accept the idea that the error terms are normally distributed.

Let's fit OLS on an R datasets and then analyze the resulting QQ plots.



The points approximately fall on the line on the x-axis are the theoretical quantiles of a standard normal. That is, we sort the n points, and then for each i, using the standard normal quantile function we find the $P_{\text{std norm}}(X \leq x) = \frac{i-0.5}{n}$ For this dataset, for the case of the leftmost point, we have that i = 1 and n = 50. which looks similar to where the leftmost point is on the x-axis. Intuitively, what this is saying is: we have 50 points and we want their x-values to be such that

$$P_{\text{std norm}}(X \leq x) = 0.01, 0.03, \cdots, 0.99.$$

We want our corresponding y to be $P_{\text{emp}}(Y \leq y) = 0.01, 0.03, \cdots, 0.99$, but based on the *empirical CDF* of the standardized residuals. What we see is that on the right hand side of the graph, the points lie slightly above the line. For the very right-most point, this is saying that the value x such that $P(X \leq x) = 0.99$ is larger under the empirical CDF for the standardized residuals than it is under a normal distribution. This suggests a 'fat tail' on the right hand side of the distribution.

1. Have enough data and invoke the central limit theorem. This is the simplest. If you have sufficient data and you expect that the variance of your errors (you can use residuals as a proxy) is finite, then you invoke the central limit theorem and *do nothing*. Your beta will be approximately normally distributed, which is all you need to construct confidence intervals and do hypothesis tests.
2. Bootstrapping. This is a non-parametric technique involving resampling in order to obtain statistics about one's data and construct confidence intervals.
3. Use a generalized linear model. Generalized linear models (GLMs) generalize linear regression to the setting of non-Gaussian errors. Thus if you think that your errors still come from some exponential family, you can look into GLMs.

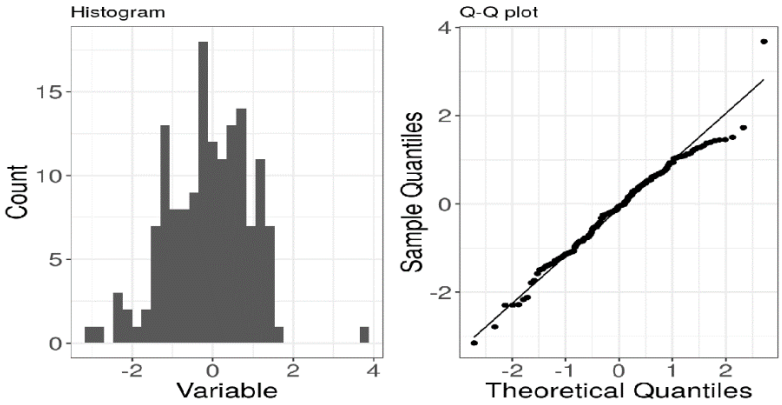
USE OF Q-Q PLOT:

To help you identify different types of distributions from a quantile-quantile plot, given examples of histograms and quantile-quantile plots for five qualitatively different distributions:

- A normal distribution
- A right-skewed distribution
- A left-skewed distribution
- An under-dispersed distribution

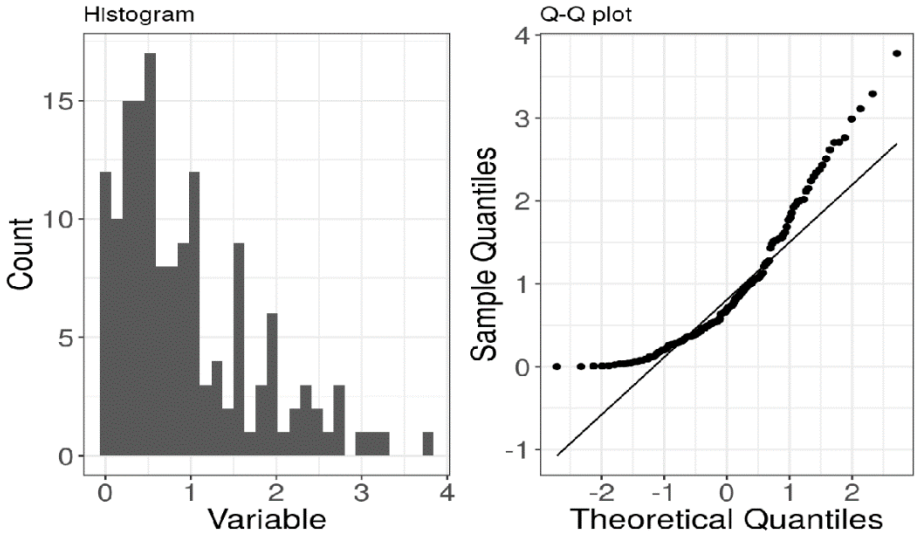
- An over-dispersed distribution

Normally distributed data:
Below is an example of data (150 observations) that are drawn from a normal distribution. The normal distribution is symmetric, so it has no skew (the mean is equal to the median). On a Q-Q plot normally distributed data appears as roughly a straight line (although the ends of the Q-Q plot often start to deviate from the straight line).



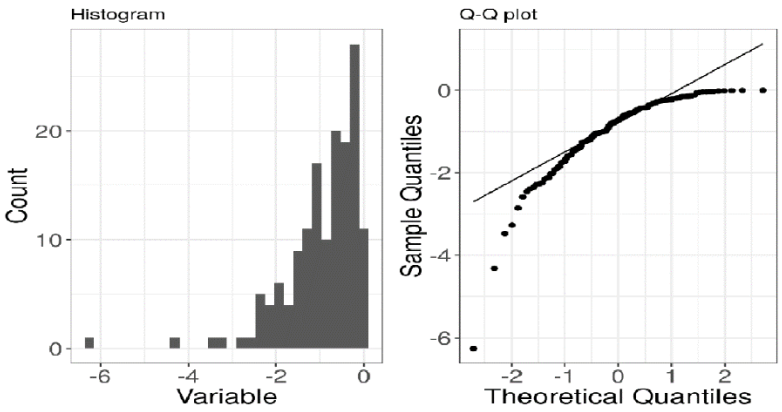
Right-skewed data

Below is an example of data (150 observations) that are drawn from a distribution that is right-skewed (in this case it is the exponential distribution). Right-skew is also known as positive skew. On a Q-Q plot right-skewed data appears curved.



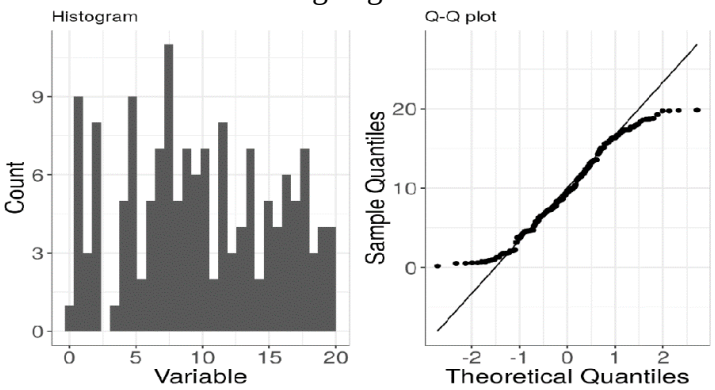
Left-skewed data

Below is an example of data (150 observations) that are drawn from a distribution that is left-skewed (in this case it is a negative exponential distribution). Left-skew is also known as negative skew.
On a Q-Q plot left-skewed data appears curved (the opposite of right-skewed data).



Under-dispersed data

Below is an example of data (150 observations) that are drawn from a distribution that is under-dispersed relative to a normal distribution (in this case it is the uniform distribution). Under-dispersed data has a reduced number of outliers (i.e. the distribution has thinner tails than a normal distribution). Under-dispersed data is also known as having a platykurtic distribution and as having negative excess kurtosis.



Over-dispersed data

Below is an example of data (150 observations) that are drawn from a distribution that is over-dispersed relative to a normal distribution (in this case it is a Laplace distribution). Over-dispersed data has an increased number of outliers (i.e. the distribution has fatter tails than a normal distribution). Over-dispersed data is also known as having a leptokurtic distribution and as having positive excess kurtosis.

On a Q-Q plot over-dispersed data appears as a flipped S shape (the opposite of under-dispersed data).