Supriya Raman
2nd December, 2019

Contents

1. Explain the linear regression algorithm in detail. 1
2. What are the assumptions of linear regression regarding residuals? 2
3. What is the coefficient of correlation and the coefficient of determination? 4
4. Explain the Anscombe's quartet in detail. 5
5. What is Pearson's R? 7
6. What is scaling? Why is scaling performed? Explain difference between normalized scaling & standardized scaling? 8
7. You might have observed that sometimes the value of VIF is infinite. Why does this happen? 9
8. What is the Gauss-Markov theorem? 10
9. Explain the gradient descent algorithm in detail. 11
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. 12

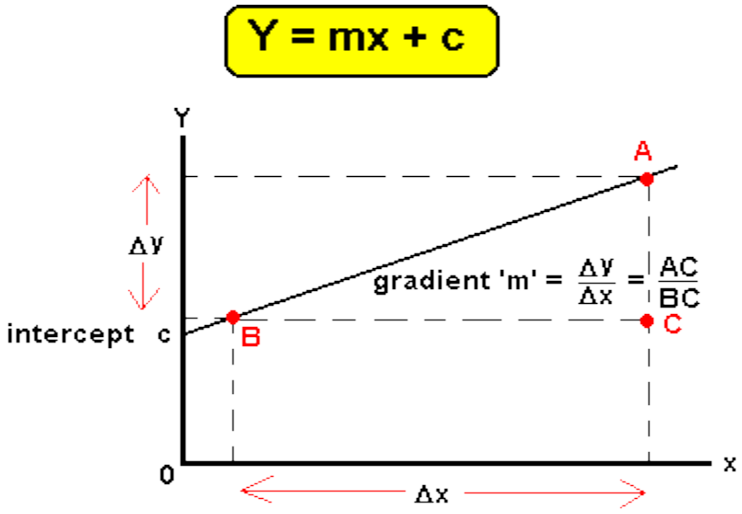
1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

Simple regression

Simple linear regression uses traditional slope-intercept form, where m and c are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction or dependent variable.

Y = mx + c



Multivariable regression

Supriya Raman
2nd December, 2019

A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

f(x, y, z) = w1x + w2y + w3z

The variables x, y and z represent the attributes or independent variables, or distinct pieces of information, we have about each observation. Example: For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

Sales=w1Radio+w2TV+w3News

Finding slope and intercept 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 m and b values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

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

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y). We reduce the Cost function or try to minimize the RMSE value to achieve the best linear line. The idea is to start with random m and b values and then iteratively updating the values, reaching minimum cost.

Regression line minimizes the sum of "Square of Residuals". That's why the method of Linear Regression is known as "Ordinary Least Square (OLS)". The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. The error is the distance between the point to the regression line.

---

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

Simple linear regression can be defined precisely as following –

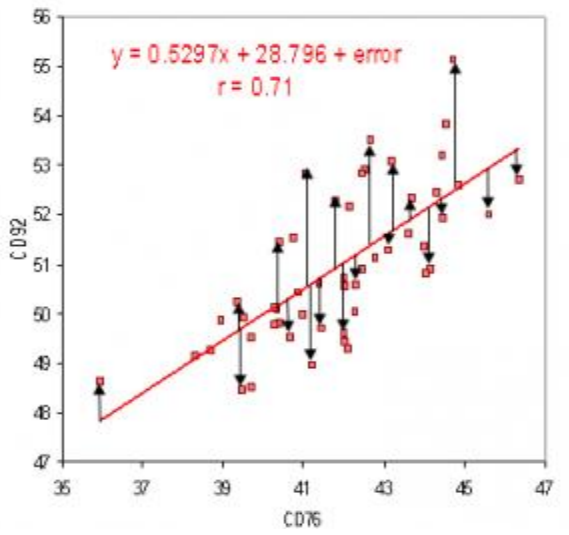
$Y_i = m * X_i + c + Error_i$

where Y_i and X_i represent the i^{th} observations of the variables Y and X respectively, m and c are fixed (but unknown) parameters and $Error_i$ is a random variable.

When you perform simple linear regression (or any other type of regression analysis), you get a line of best fit. The data points usually don't fall exactly on this regression equation line; they are scattered around. A residual is the vertical distance between a data point and the regression line. Each data point has one residual. They are positive if they are above the regression line and negative if they are below the regression line. If the regression line actually passes through the point, the residual at that point is zero.

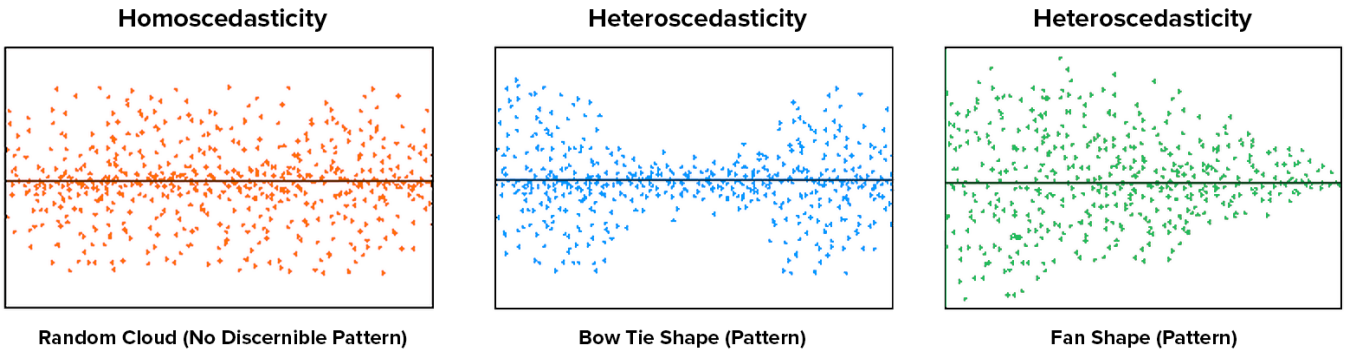
Error values in below graph is shown by arrows –

Supriya Raman
2nd December, 2019



Assumptions of linear regression regarding residuals are as following: -

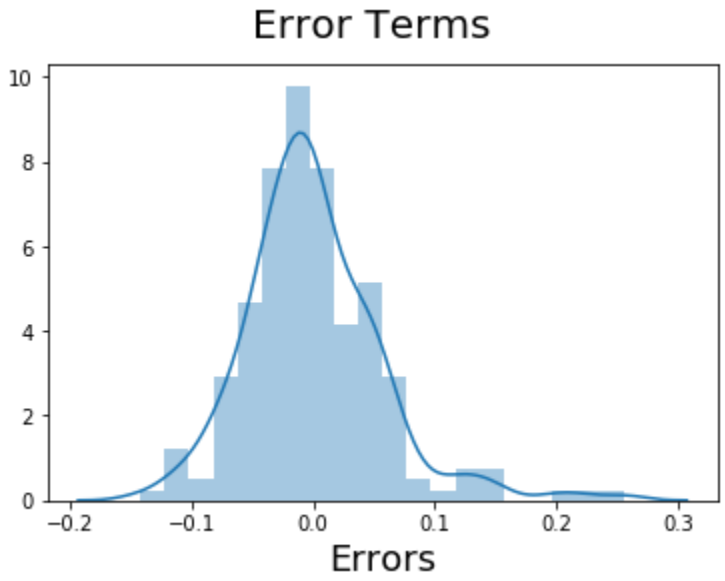
1. Homoscedasticity of residuals or equal variance: - Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a good way to check for homoscedasticity. There should be no clear pattern in the distribution and if there is a specific pattern, the data is heteroscedastic.



The leftmost graph shows no definite pattern i.e. constant variance among the residuals, the middle graph shows a specific pattern where the error increases and then decreases with the predicted values violating the constant variance rule and the rightmost graph also exhibits a specific pattern where the error decreases with the predicted values depicting heteroscedasticity.

2. Normal distribution of error terms – The error terms when plotted on a distribution plot or histogram should represent a normal distribution graph which has mean of residuals is zero or something close to zero and error terms are having equal variance thus plotting a bell shaped curve as below -

Supriya Raman
2nd December, 2019



3. The X variables and residuals are uncorrelated – The predictor values and error term should not be correlated. We can do a correlation test on the X variable and the residuals. If p-value is high, so null hypothesis that true correlation is 0 can't be rejected. So, the assumption holds true for this model.

4. No Multicollineraity - Multicollinearity refers to the phenomenon of having related predictor variables in the input dataset. In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, due to which the presence of that variable in the model is redundant. We should drop some of these related independent variables as a way of dealing with multicollinearity.

   Multicollineraity doesn't affect the final prediction or precision of prediction of the model. However, the interpretation and inference of the model does change. It is essential to detect and deal with the multicollinearity present in the model. WE can detect the multicollinearity by following two ways –
   a) Looking at pairwise correlations between different pairs of independent variables – This can be achieved by ploting a heatmap or a pair plot.
   b) Checking the Variance Inflation Factor (VIF) as instead of just one variable, the independent variable might depend upon a combination of other variables. The VIF is given by:

      $VIF_i = 1 / (1-R_i^2)$

      If VIF value is greater than 10 then variable should be eliminated from model. VIF greater than 5 can be okay, but it is worth inspecting. VIF less than 5 is good VIF value and variables must not be eliminated.

---

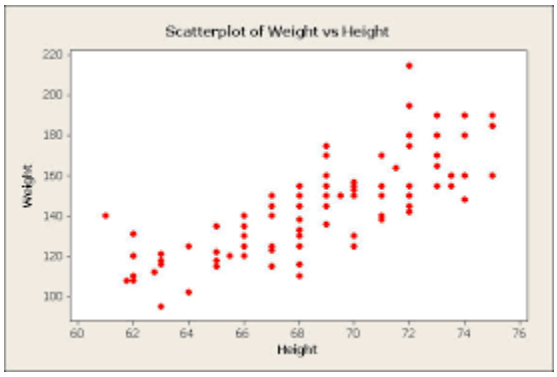
3. What is the coefficient of correlation and the coefficient of determination?

#	Co-efficient of Correlation	Co-efficient of Determination
1.	"R" value which is given in the summary table in the Regression output	R-squared value = When we multiply R times R, we get the R square value.
2.	Square root of Co-efficient of Determination	Square of Coefficient of Correlation
3.	Measures linear relationship between two variables or expresses the presence or non-presence of a linear interrelationship between the two observed variables.	Measures explained variation or implied causation.

Supriya Raman
2nd December, 2019

4.	Degree of relationship between two variables say x and y.	Percentage variation in y which is explained by all the x variables together.
5.	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.	Higher the better. It's value is always between 0 and 1. It can never be negative – since it is a squared value.
6.	Correlation can be rightfully explained for simple linear regression – because we 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. R square is for both simple linear regressions and also for multiple linear regressions

Example: Height and weight of individuals are correlated. If the correlation coefficient is r = 0.8 it signifies there is high positive correlation. Both height and weight of individuals' increase/decrease together (positive) and their relationship (linear) is strong. The scatter plot is something like below –



But height of individuals may also be affected by other factors like age, genetics, food intake, amount and type of exercise, location etc. So, we if try to predict height by using weight as a single predictor, coefficient of determination is 0.64 (equals to square of correlation coefficient here). It shows that 0.64 (or 64%) of variation in height can be explained by weight. and remaining 36% of variation in height may be due to other factors which affect height of individuals like age, genetics, food intake, amount and type of exercise, location etc.

---

4. Explain the Anscombe's quartet in detail.

People tended to ignore visualizations in favor of summary statistics. Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimize against and use as a barometer for our business. But there's a danger in relying only on summary statistics and ignoring the overall distribution. We find its too much effort to plot the data.

Anscombe's quartet is constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. It comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

Linear Regression Assignment

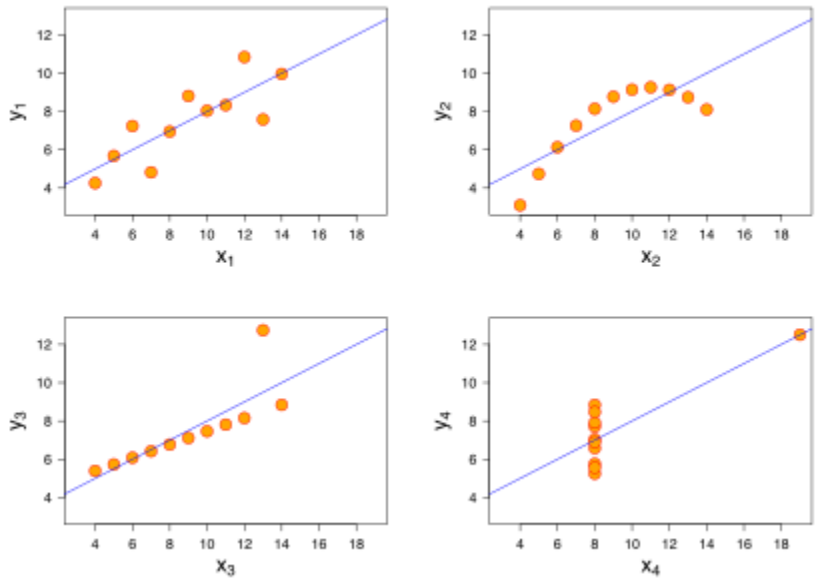
Supriya Raman
2nd December, 2019

Below four quartet shows same descriptive statistics. The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

	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

But things change completely when they are graphed.



- The first scatter plot (top left) appeared to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.

Supriya Raman
2nd December, 2019

- The second graph (top right) was not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

---

5. What is Pearson's R?

Pearson's correlation coefficient or Pearson's R 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.

Pearson's correlation coefficient formula for two variables x and y is as below where

- n is sample size
- x and y are the individual sample points

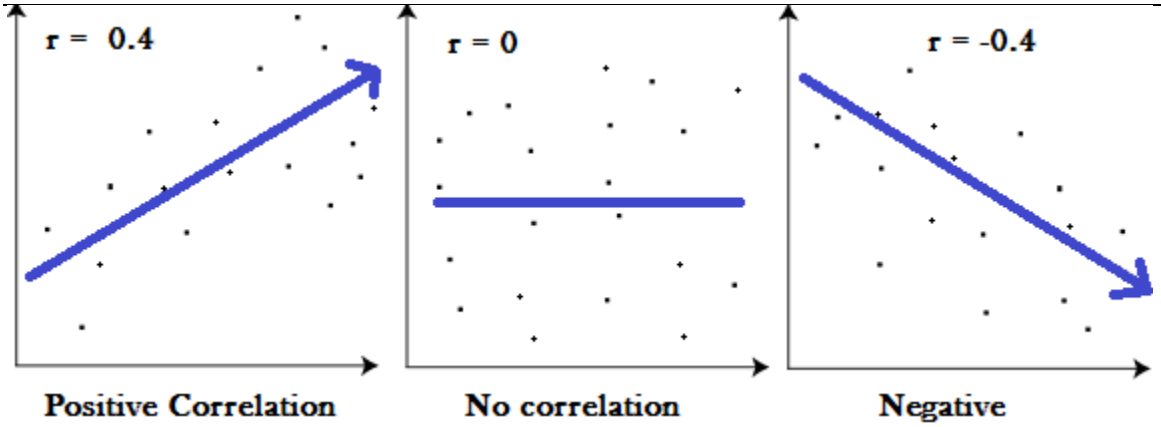
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Degree of correlation:

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1.

- Positive correlation indicates every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
  Example: shoe sizes go up in (almost) perfect correlation with foot length.
- Negative correlation coefficient means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
  Example: the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- A result of zero indicates no relationship at all. The two variables just aren't related.
- +1 is a perfect positive relationship. -1 is a perfect negative relationship.

Supriya Raman
2nd December, 2019



6. What is scaling? Why is scaling performed? Explain difference between normalized scaling & standardized scaling?

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.

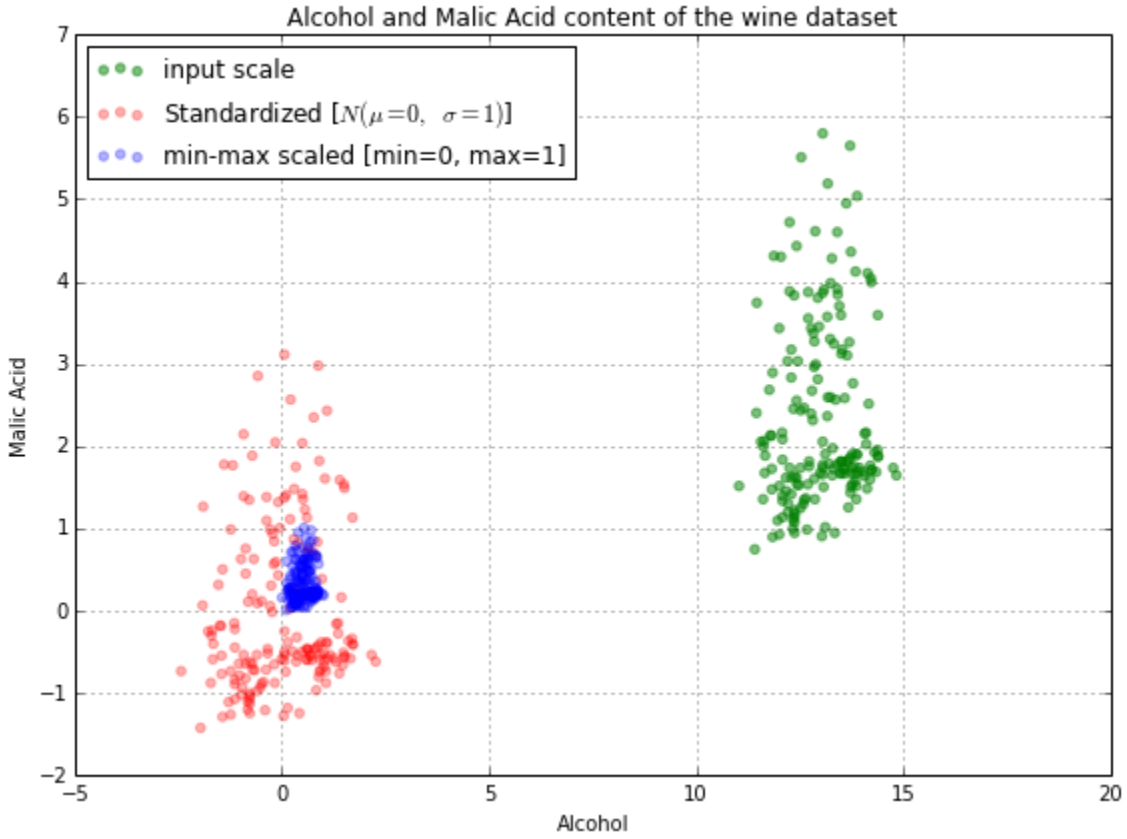
Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Two common methods to perform Feature Scaling are as below –

#	Standardization Scaling	Normalized Scaling or Min-Max Scaling
1.	It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1	This technique re-scales a feature or observation value with distribution value between 0 and 1
2.	$x' = \frac{x - \bar{x}}{\sigma}$	$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$
3.	This distribution will have values between different ranges but they are centered around 0 with a standard deviation of 1	This distribution will have values between 0 and 1 .
4.	Outlier values are also just Standardized but not removed.	It removes outliers in the sample

Example: Below graph represents the actual data points and Standardized and Min-Max Scaled data points.

Supriya Raman
2nd December, 2019



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

An infinite VIF (Variance Inflation Factors) value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

VIF of a variable X1 = 1 / (1 - R-squared of X1 on all other X's)

VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. The VIF can be conceived as related to the R-squared of a particular predictor variable regressed on all other includes predictor variables.:

If we only have one X or that X is orthogonal with all the other X values; then VIF = 1 / (1-0) = 1 In this case, there is no variance inflation.

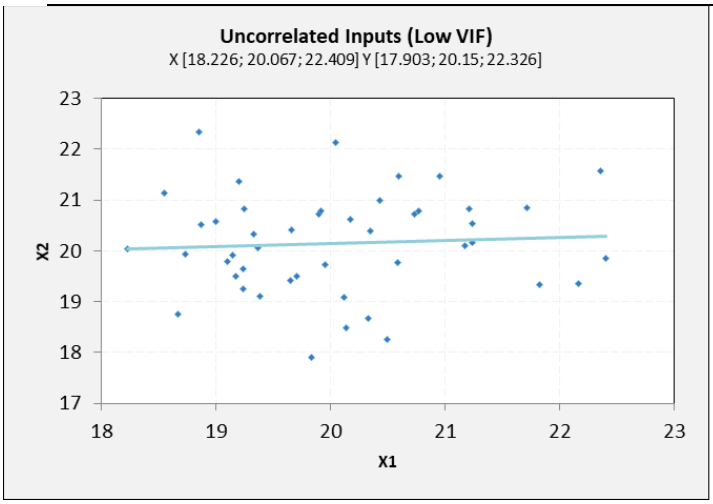
If two X values are perfectly correlated then R-squared value is 1, VIF = 1 / (1-1) = 1/0 = Infinity.

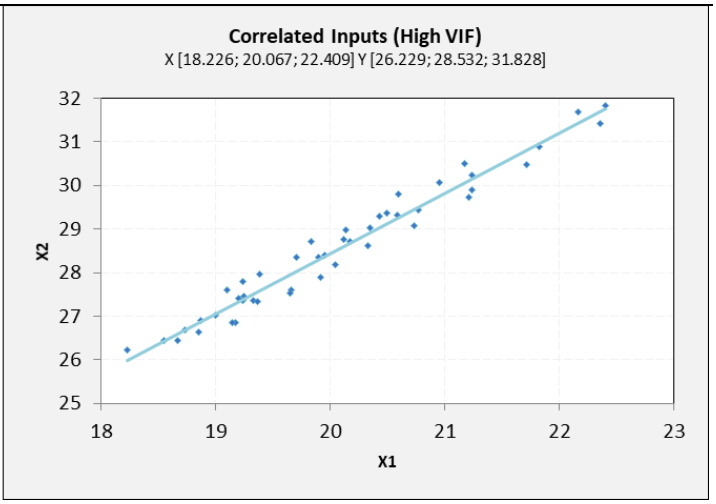
When building a multiple regression model, we should always check VIF values for independent variables and determine if we need to take any corrective action before building the model. Following are some of the corrective actions –

1. We should review independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF

for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then we may need to drop more terms as required.

2. A second approach is to use principal component analysis and determine the optimal set of principal components that best describe independent variables. Using this approach will get rid of multicollinearity problem but it may be hard to interpret the meaning of these "new" independent variables.
3. The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
4. The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
5. Finally, we can use a different type of model call ridge regression that better handles multicollinearity.





8. What is the 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 possible with smallest variance of all possible linear estimators.

The Gauss-Markov theorem famously states that OLS is BLUE. BLUE is an acronym for Best Linear Unbiased Estimator.

- Best signifies that variance of OLS estimator is minimal, smaller than variance of any other estimator.
- Linear signifies that if relationship is not linear, OLS is not applicable.
- Unbiased say that expected values of estimated beta and alpha equal true values describing relationship of x and y.

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

1. Linearity: The parameters we are estimating using the OLS method must be themselves linear. The value of y can be linearly expressed in terms of x

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

   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.

2. Random: Our data must have been randomly sampled from the population. This signifies that expected value of error term is zero for all the observations.

$E\{\varepsilon i\} = 0, i = 1, \ldots, N$

3. Homoscedasticity: No matter what the values of our regressors might be, the error of the variance is constant. The conditional variance of the error term is constant for all x values and over time.
   $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2 = \text{constant}$

4. Error term is independently distributed and not correlated, no correlation between observations.
   $cov\{\varepsilon i, \varepsilon j\} = E\{\varepsilon i, \varepsilon j\} = 0, i, j = 1,\ldots., N \text{ and } i \neq j.$
   {ε1……εn} and {x1…..,xN} are independent

5. Xi is deterministic. X is uncorrelated with error term since xi is deterministic.
   $Cov(Xi, \varepsilon i) = E(Xi\ \varepsilon i) – E(Xi) * E(\varepsilon i) = Xi*E(\varepsilon i) – Xi * E(\varepsilon i) = 0$ because Xi is deterministic

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 we know where these conditions are violated, we may be able to plan ways to change your experiment setup to help our situation fit the ideal Gauss Markov situation more closely.

---

9. Explain the gradient descent algorithm in detail.

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.

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 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.

A Loss Functions (Cost Function) 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.

<u>Simple Regression</u>
Given our simple linear equation y = mx + b; we can calculate Mean Squared Error as:

$$f(m,b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Where

- N is the total number of observations (data points)
- y_i is the actual value of an observation and $mx_i + b$ is our prediction

There are two parameters (coefficients) in our cost function we can control: weight m and bias b. Since we need to consider the impact each one has on the final prediction, we use partial derivatives. To find the partial derivatives, we use the Chain rule. We need the chain rule because $(y–(mx+b))^2$ is really 2 nested functions: the inner function y–(mx+b) and the outer function x^2. We can calculate the gradient of this cost function as:

Supriya Raman
2nd December, 2019

$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -x_i \cdot 2(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -1 \cdot 2(y_i - (mx_i + b)) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

Multiple Linear Regression

$y = W_1x_1 + W_2x_2 + W_3x_3$; we can calculate Mean Squared Error as below:

$$\frac{1}{2N}\sum_{i=1}^{n}(y_i - (W_1x_1 + W_2x_2 + W_3x_3))^2$$

Again using the Chain rule, we can compute the gradient–a vector of partial derivatives describing the slope of the cost function for each weight.

$$f'(W_1) = -x_1(y - (W_1x_1 + W_2x_2 + W_3x_3))$$
$$f'(W_2) = -x_2(y - (W_1x_1 + W_2x_2 + W_3x_3))$$
$$f'(W_3) = -x_3(y - (W_1x_1 + W_2x_2 + W_3x_3))$$

---

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

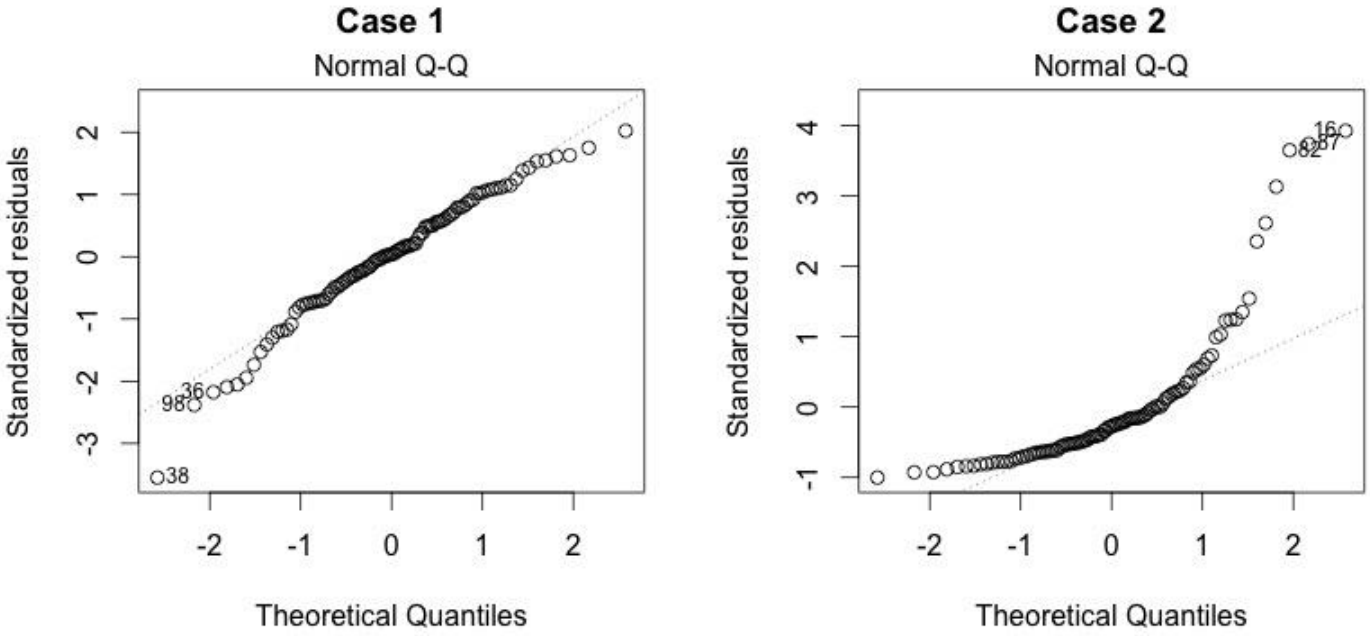
Q-Q is Quantile-Quantile plot. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

It is an exploratory graphical device used to check the validity of a distributional assumption for a data set. We compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a straight line.

The definition of the q-q plot may be extended to any continuous density (Normal Q-Q Plot). The q-q plot will be close to a straight line if the assumed density is correct. Because the cumulative distribution function of the uniform density was a straight line, the q-q plot was very easy to construct. For data that are not uniform, the theoretical quantiles must be computed in a different manner.

Normal Q-Q Plot is used to assess if your residuals are normally distributed. Residuals in Linear Regression shows the predicted value (based on the regression equation) on the X axis, and the residuals on the Y axis. The residuals are essentially the difference between the predicted value and the actual value (i.e. the 'error' in your predicted value). If Q-Q Plot of Residuals has the data points closely following the straight line at a 45% angle upwards (left to right), it shows that residuals are normally distributed.

Supriya Raman
2nd December, 2019



Above graph – Case 1 shows quantiles from the standard Normal distribution with mean 0 and standard deviation 1. Case 2 shows a curve instead of a straight line. Normal Q-Q plots that look like this usually mean sample data are skewed.