Subjective Questions

Question 1 - Explain the linear regression algorithm in detail.

Linear Regression

Linear regression is a type of supervised learning algorithm, commonly used for predictive analysis.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called **the dependent variable.** The variable you are using to predict the other variable's value is called **the independent variable**.

This form of analysis estimates the coefficients of the linear equation, **involving one or more independent variables that best predict the value of the dependent variable**. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a **"least squares" method to discover the best-fit line for a set of paired data.** You then estimate the value of X (dependent variable) from Y (independent variable).

Linear Regression is classified based on no. of independent variables

1. Simple linear Regression
2. Multiple linear Regression

Simple Linear Regression

The most elementary type of regression model is the simple linear regression which explains the **relationship between a dependent variable and one independent variable** using a straight line. The straight line is plotted on the scatter plot of these two points.

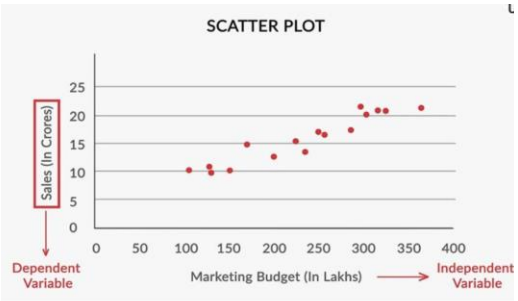


Figure 2 – Scatter plot

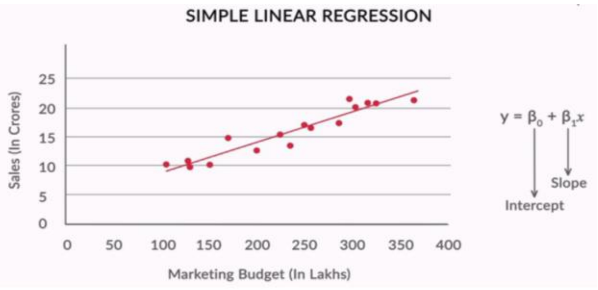


Figure 3 - Regression Line

The standard equation of the regression line is given by the following expression:–

$$Y = \beta_0 + \beta_1 * X$$

It is used in estimating exactly how much of **Y will change, when x changes a certain amount.**

Multiple Linear Regression

Multiple linear regression is a statistical technique to understand the **relationship between one dependent variable and several independent variables (explanatory variables).**The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Consider our previous **example of sales prediction** using TV Marketing budget. In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e. TV marketing, Radio marketing, and Newspaper marketing.

Multiple linear regression model is also built on a straight line. The equation of multiple linear regression would be as follows:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 \text{ ….} + \beta_n * X_n$$

The significance of this model lies in the fact that one can easily interpret and understand the marginal changes and their consequences. For example, if the value of x_0 increases by 1 unit, keeping other variables constant, the total increase in the value of *y* will be β_i. Mathematically, the intercept term (β_0) is the response when all the predictor terms are set to zero or not considered.

For example of sales prediction using TV, Radio, and Newspaper marketing will be

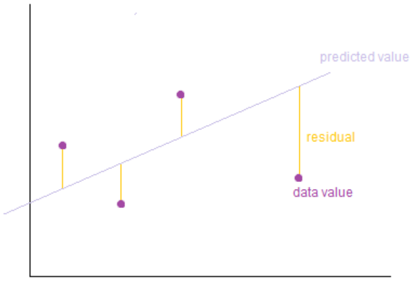
$$Y = \beta_0 + \beta_1 * \text{T.V. marketing} + \beta_2 * \text{Internet marketing} + \beta_3 * \text{Newspaper marketing}$$

You built the model containing all variables in python using **sklearn.linear_model**

Best Fit line

The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

When you perform linear regression you get a line of best fit. As shown below.



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.

Residuals

As residuals are the difference between any data point and the regression line, they are sometimes called "**errors**." Error in this context doesn't mean that there's something wrong with the analysis; it just means that there is some unexplained difference. In other words, the residual is the error that isn't explained by the regression line.

The residual **(e)** can also be expressed with an equation. The e is the difference between the predicted value **(y_{pred})** and the observed value **(y_i)**. The scatter plot is a set of data points that are observed, while the regression line is the prediction.

Residual = Observed value – predicted value

$$e = y_i - y_{pred}$$

Strength of Linear Regression

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

1. **R^2 or Coefficient of Determination**

   R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes.

   Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as:

$$\mathbf{R^2 = 1 -(RSS / TSS)}$$

where:

RSS IS Residual sum of square
TSS is sum of errors of the data from mean

RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

$$\boxed{RSS = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2}$$

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$\mathbf{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Assumptions in linear regression Model.

1. **Assumption about the form of the model:**
   It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.
2. **Assumptions about the residuals:**
   a. **Normality assumption:** It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.
   b. **Zero mean assumption**: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
   c. **Constant variance assumption**: It is assumed that the residual terms have the same (but unknown) variance, σ^2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
   d. **Independent error assumption**: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.
3. **Assumptions about the estimators:**
   a. The independent variables are measured without error.
   b. The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.

Importance of Linear regression

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

Few real time examples

1. **Evaluating trends and sales estimates**
   You can also use linear-regression analysis to try to predict a salesperson's total yearly sales (the dependent variable) from independent variables such as age, education and years of experience
2. **Analyze pricing elasticity**
   Changes in pricing often impact consumer behavior — and linear regression can help you analyze how. For instance, if the price of a particular product keeps changing, you can use regression analysis to see whether consumption drops as the price increases. What if consumption does not drop significantly as the price increases? At what price point do buyers stop purchasing the product? This information would be very helpful for leaders in a retail business.

Question 2 - What are the assumptions of linear regression regarding residuals?

Assumptions about the residuals:

1. **Normality assumption:** It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.
2. **Zero mean assumption**: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
3. **Constant variance assumption**: It is assumed that the residual terms have the same (but unknown) variance, σ^2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
4. **Independent error assumption**: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.
5. **Lack of perfect multicollinearity in the predictors**. For standard least squares estimation methods, the design matrix X must have full column rank p; otherwise, we have a condition known as perfect multicollinearity in the predictor variables. This can be triggered by having two or more perfectly correlated predictor variables

Question 3 - What is the coefficient of correlation and the coefficient of determination?

Coefficient of correlation

Coefficient of correlation is "R" value which is given in the summary table in the Regression output. R^2 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.

1. **R^2 or Coefficient of Determination**

   It 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 a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

   Mathematically, it is represented as:

   $$\mathbf{R^2 = 1 -(RSS / TSS)}$$

   where:

   RSS IS Residual sum of square
   TSS is sum of errors of the data from mean

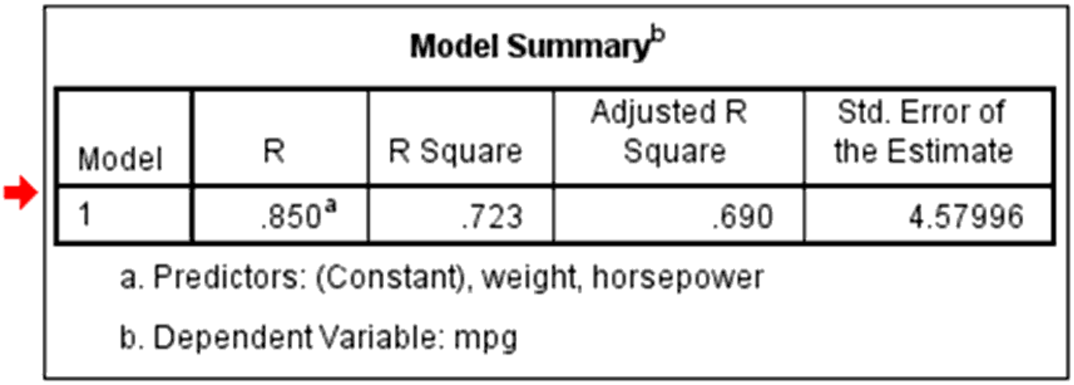
   RSS (Residual Sum of Squares): In statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

   $$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

The correlation coefficient expresses the presence or non-presence of a linear interrelationship between the two observed variables. If the linear interrelationship is positive, the correlation coefficient will be a positive number between 0 and 1.0. If, on the other hand, it is negative, the number will be between 0 and -1.0.

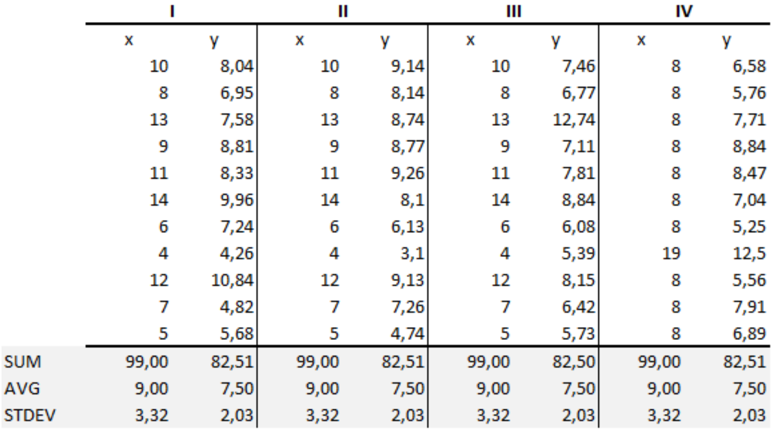
In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data. Values of R^2 outside the range 0 to 1 can occur when the model fits the data worse than a horizontal hyperplane. R^2 is a statistic that will give some information about the ***goodness of fit*** of a model.

2. **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^2 is a better term. You can explain R^2 for both simple linear regressions and also for multiple linear regressions.

Question 4 Explain the Anscombe's quartet in detail

Anscombe's quartet

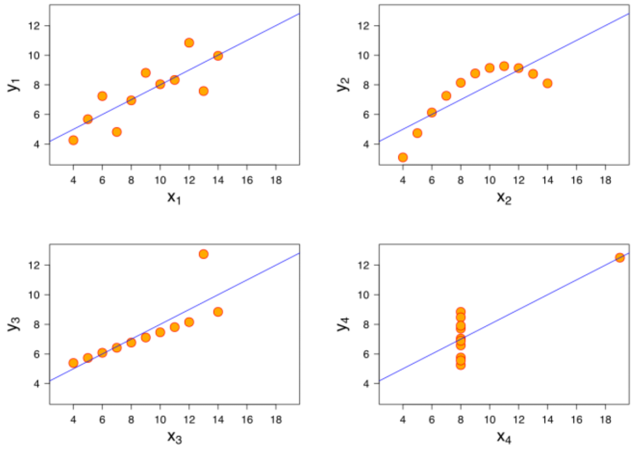
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



	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

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
- When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

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. Data visualization is crucial in developing a sensible statistical model. **"A computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.**"

Significance of Anscombe's Quartet

1. Anscombe's Quartet is a great demonstration of the **importance of graphing data to analyze it**. Given simply variance values, means, and even linear regressions cannot accurately portray data in its native form. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.
2. Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict.

Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

Question 5 - What is Pearson's R?

Pearson's R

Also referred as **the Pearson correlation coefficient** (PCC), the **Pearson product-moment correlation coefficient** (PPMCC) or the **bivariate correlation**

The Pearson R is a measure of the strength of the linear relationship between two variables. Basically, it attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit). If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

The symbol for Pearson's correlation is **"ρ"** when it is measured in the population and **"r"** when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use **"r"** to represent Pearson's correlation unless otherwise noted.

Given a pair of random variable (X, Y), the formula for p is

$$\rho = \frac{cov\ (X,Y)}{\sigma x \sigma y}$$

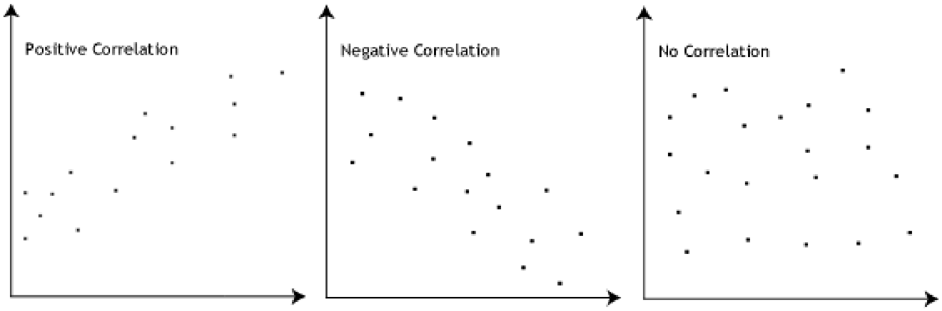
where:

- cov - is the covariance
- σ_x is standard deviation of X
- σ_y is standard deviation of Y

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1.

1. A value of 0 indicates that there is no association between the two variables.
2. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
3. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below:



Guidelines to interpreting Pearson's correlation coefficient

	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Remember that these values are guidelines and whether an association is strong or not will also depend on what you are measuring

Key Points -

1. **Type of Variable -** Pearson's correlation coefficient **doesn't depend on type of the variable**. Two variables have to be measured on either an interval or ratio scale. However, both variables do not need to be measured on the same scale (e.g., one variable can be ratio and one can be interval)

2. **Different Units - Two variables can be measured in entirely different units.** For example, you could correlate a person's age with their blood sugar levels. Here, the units are completely different; age is measured in years and blood sugar level measured in mmol/L (a measure of concentration). Indeed, the calculations for Pearson's correlation coefficient were designed such that the units of measurement do not affect the calculation. This allows the correlation coefficient to be comparable and not influenced by the units of the variables used.

3. **Dependent and Independent variable** - **The Pearson R does not take into consideration whether a variable has been classified as a dependent or independent variable**. It treats all variables equally. For example, you might want to find out whether basketball performance is correlated to a person's height. You might, therefore, plot a graph of performance against height and calculate the Pearson correlation coefficient.

4. The Pearson correlation coefficient, *r*, **does not represent the slope of the line of best fit.** Therefore, if you get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another. It simply means that there is no variation between the data points and the line of best fit.

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

What is scaling?

Feature scaling is a method used to normalize the range of independent variables or **features** of data. It is also known as data normalization. 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. It is generally performed during the data preprocessing step.

Feature scaling is one of the main components of data preprocessing, and can be applied to all types of data one might come across.

Common Methods to perform scaling are –

1. Standardization (mean-0,sigma -1)
2. Mean Normalization (normalization between -1 and 1)
3. Min-Max scaling (normalization): Between 0 and 1.
4. Unit Vector

I. **Standardization**

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:

$$z = \frac{x - \mu}{\sigma}$$

sklearn.preprocessing.scale helps us implementing standardization in python.

II. **Mean Normalization**

This distribution will have values between **-1 and 1with μ=0.**

$$X' = \frac{\boldsymbol{x - mean(x)}}{\boldsymbol{max(x) - min(x)}}$$

Standardization and **Mean Normalization** can be used for algorithms that assumes zero centric data like **Principal Component Analysis (PCA).** This is another form of the standardization scaling.

III. **Min Max scaling**

$$X' = \frac{\boldsymbol{x - min(x)}}{\boldsymbol{max(x) - min(x)}}$$

This scaling brings the value between 0 and 1.

IV. **Scaling to unit length (Unit Vector)**

Another option that is widely used in machine-learning is to scale the components of a feature vector such that the complete vector has length one. This usually means dividing each component by the **Euclidean length of the vector**. In other words, scaling is done considering the whole feature vector to be of unit length.

$$X' = \frac{x}{||x||}$$

Min-Max Scaling and **Unit Vector** techniques produces values of range [0, 1]. When dealing with features with hard boundaries this is quite useful. For example, when dealing with image data, the colors can range from only 0 to 255.

Why is scaling performed?

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.

Simplest way to explain **why feature scaling is important** is -

Consider k-means clustering. If your data has 4 columns, with 3 of the columns' values being scaled between 0 and 1, and the fourth possessing values between 0 and 1,000, it is relatively easy to determine which of these features will be the sole determinant of clustering results.

Advantages of scaling

1. Interpretability becomes easier.
2. Gradient descent converges much faster with feature scaling than without it. "Normalized" data set tends to require less number of steps to reach the minimum of the cost function compared to without scaled data.
3. Standardizing the features so that they are centered around 0 with a standard deviation of 1 helps in comparing measurements that have different units
4. k-nearest neighbors with a Euclidean distance measure is sensitive to magnitudes and hence should be scaled for all features to weigh in equally.
5. Scaling is critical, while performing Principal Component Analysis (PCA). PCA tries to get the features with maximum variance and the variance is high for high magnitude features. This skews the PCA towards high magnitude features

What is the difference between normalized scaling and standardized scaling?

Normalization vs. Standardization

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:

$$z = \frac{x - \mu}{\sigma}$$

The advantage of **Standardization** over the other is that it doesn't compress the data between a particular range as in **Min-Max scaling**. This is useful, especially if there is are extreme data point (outlier)

In statistics, Standardization is the subtraction of the mean and then dividing by its standard deviation. In Algebra, Normalization is the process of dividing of a vector by its length and it transforms your data into a range between 0 and 1.

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

Brief bout VIF (Variance Inflation Factor)

A variance inflation factor (VIF) detects multicollinearity in regression analysis. Multicollinearity is when there’s correlation between predictors (i.e. independent variables) in a model; its presence can adversely affect your regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

In statistics, the variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. “i” is the predictor you’re looking at (e.g. x1 or x2):

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

Interpreting the Variance Inflation Factor

Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors. A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

VIF = infinity indicates that predictor variable is having perfect correlation with other predictors.

Question 8 - What is the Gauss-Markov theorem?

Gauss-Markov

In statistics, **the Gauss–Markov theorem states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectation value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists**. Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance.

When estimating regression models, we know that the results of the estimation procedure are random. However, when using unbiased estimators, at least on average, we estimate the true parameter. When comparing different unbiased estimators, it is therefore interesting to know which one has the highest precision: being aware that the likelihood of estimating the exact value of the parameter of interest is 0 in an empirical application, we want to make sure that the likelihood of obtaining an estimate very close to the true value is as high as possible. This means we want to use the estimator with the lowest variance of all unbiased estimators, provided we care about unbiasedness. The Gauss-Markov theorem states that, in the class of conditionally unbiased linear estimators, the OLS estimator has this property under certain conditions.

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

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

The Gauss-Markov Assumptions In Algebra

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

$$\mathbf{Y_i = x_i' \beta + \varepsilon_i}$$

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

$E\{\varepsilon_i\} = 0, i = 1, . . . , N$
$\{\varepsilon_1 \ldots\ldots \varepsilon_n \}$ & $\{x_1 \ldots\ldots, x_N \}$ are independent
$cov\{\varepsilon_i, \varepsilon_j\} = 0, i, j = 1, \ldots., N \; I \neq j.$
$V\{\varepsilon_1 = \sigma^2, i= 1, \ldots.N$

The first of these assumptions can be read as "**The expected value of the error term is zero.".** The second assumption is **collinearity**, the third is **exogeneity**, and the fourth is **homoscedasticity**.

The Gauss-Markov Theorem for $\hat{\beta}_1$

Gauss-Markov TheoremIThe theorem states that $\boldsymbol{\beta_1}$ has minimum variance among allunbiased linear estimators of the form

$$\hat{\beta}1=\sum ciYi$$

As this estimator must be unbiased we have

$$E\{\hat{\beta}_1\}=\sum c_iE\{Y_i\}=\beta_1=\sum c_i(\beta_0+\beta_1X_i) =\beta_0\sum c_i+\beta_1\sum c_iX_i=\beta_1$$

This imposes some restrictions on the ci's.

Given these constraints:

$$\beta_0\sum c_i+\beta_1\sum c_iX_i=\beta_1$$

clearly it must be the case that

$$\sum c_i= 0 \text{ and} \sum c_iX_i= 1$$

The variance of this estimator is

$$\sigma^2\{\hat{\beta}_1\}=\sum c_i^2\sigma^2\{Y_i\}=\sigma^2\sum c_i^2$$

This also places a kind of constraint on the c_i's

Now define $c_i=k_i+d_i$

where k_i are the constants we already defined and d_i are arbitrary constants.

Let's look at thevariance of the estimator

$$\sigma^2\{\hat{\beta}_1\}=\sum c_i^2\sigma^2\{Y_i\}=\sigma^2\sum(k_i+d_i)^2=\sigma^2(\sum k_i^2+\sum d_i^2+ 2\sum k_id_i)$$

Note we just demonstrated that

$$\sigma^2\sum k_i^2=\sigma^2\{b_1\}$$

So

$\sigma^2\{\hat{\beta}_1\}$ is related to $\sigma^2\{b_1\}$ plus some extra stuff.

Now by showing that $\sum k_id_i= 0$

we're almost done.

$$\sum k_id_i=\sum k_i(c_i-k_i)=\sum k_i(c_i-k_i)=\sum k_ic_i-\sum k_i^2=\sum c_i(X_i-\bar{X}\sum(X_i-\bar{X})^2)-1\sum(Xi-\bar{X})^2=\sum c_iX_i-\bar{X}\sum c_i\sum(X_i-\bar{X})^2-1\sum(X_i-\bar{X})^2= 0$$

So we are left with $\sigma^2\{\hat{\beta}_1\}=\sigma^2(\sum k_i^2+\sum d_i^2)=\sigma^2(b_1) +\sigma^2(\sum d_{i2})$

which is minimized when the $d_i= 0\ \forall$ i.

If $d_i= 0$ then $c_i=k_i$.

This means that the least squares estimator b1 has minimum variance among all unbiased linear estimators

Question 9 - Explain the gradient descent algorithm in detail.

Gradient Descent Algorithm

Gradient descent is an optimization algorithm. In linear regression, it is used to optimize the cost function and find the values of the βs (estimators) corresponding to the optimized value of the cost function.

Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).

Or

A person is stuck in the mountains and is trying to get down (i.e. trying to find the global minimum). the path down the mountain is not visible. They can use the method of gradient descent, which involves looking at the steepness of the hill at their current position, then proceeding in the direction with the steepest descent (i.e. downhill). Using this method, they would eventually find their way down the mountain (i.e. local minimum.

However, assume also that the steepness of the hill is not immediately obvious with simple observation, but rather it requires a sophisticated instrument to measure, which the person happens to have at the moment. It takes quite some time to measure the steepness of the hill with the instrument, thus they should minimize their use of the instrument if they wanted to get down the mountain before sunset. The difficulty then is choosing the frequency at which they should measure the steepness of the hill so not to go off track.

In this analogy, the person represents the algorithm, and the path taken down the mountain represents the sequence of parameter settings that the algorithm will explore. The steepness of the hill represents the **slope of the error surface** at that point. The instrument used to measure steepness is differentiation **(the slope of the error surface can be calculated by taking the derivative of the squared error function at that point)**. The direction they choose to travel in aligns with the gradient of the **error surface at that point**. The amount of time they travel before taking another measurement is the **learning rate of the algorithm**

Learning rate

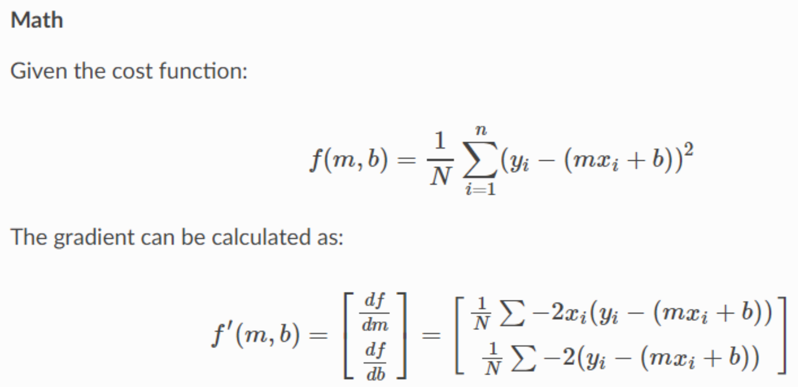
The size of these steps is called the *learning rate*. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient 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 Cost 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: mm (weight) and bb (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.



Math

Given the cost function:

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

The gradient can be calculated as:

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

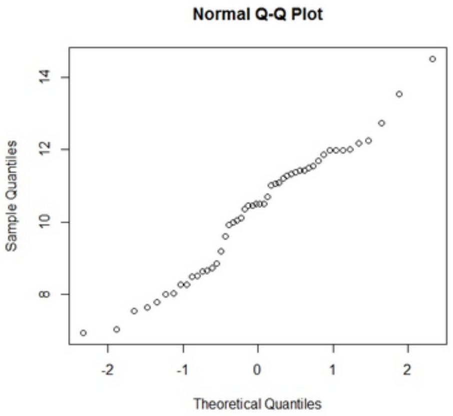
To solve for the gradient, we iterate through our data points using our new mm and bb values and compute the partial derivatives. This new gradient tells us the slope of our cost function at our current position (current parameter values) and the direction we should move to update our parameters. The size of our update is controlled by the learning rate.

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

Q-Q plot

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is normally distributed, we can use a Normal Q-Q plot to check that assumption. **It's just a visual check, not an air-tight proof**, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

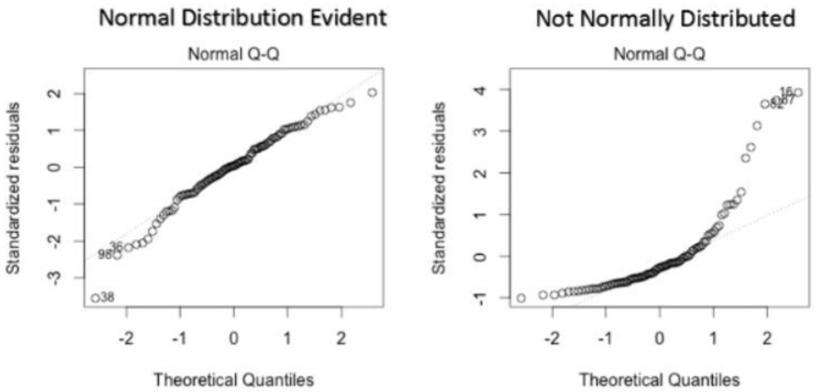
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal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



Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot help to validate the assumption of the Linear equation that error terms are normally distributed. In other words, Normal distribution of the residuals can be validated by plotting a q-q plot. Using the q-q plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line.

Most statistical procedures in inferential statistics assumes that the sample drawn from the population has a normal property. There are different ways of identifying whether a given variable has a Normal property or not. One of the way is using Q-Q plot.



Normal Distribution Evident

Not Normally Distributed

Consider a sample data from a population. The construction of Q-Q plots starts with ordering the sample data from smallest to largest. Let k denote the ranking or order number. Therefore k=1 for the smallest and k = n for the largest. The q-q plot is based on the fact that the ordered value of k is an estimate of (k-1/2)/n quantile of the sample data. In other words, **the ordered values are close to inverse of cumulative distribution of (k-0.5)/n, where n is the sample size**. If the cumulative distribution function belongs to an appropriate known distribution, **then the plot of ordered values and the known cumulative distributional values will approximately form a straight line**. On the other hand, if the assumed distribution is inappropriate, the points will deviate from the straight line in a systematic manner. **Therefore, if we assume that the cdf is from a normal distribution, then obtaining a straight line after the plot confirms that the sample data indeed belongs to the normal population.**