Linear Regression Subjective Questions Assignment

Submitted By: Mudassar Majgaonkar

1. Explain the linear regression algorithm in detail.

Ans: Regression is a method of modelling a target value based on independent input variables. Linear Regression is a most basic type of statistical model, which assumes that there is a linear relationship between the target and the independent input variables. This model gives us the strength and direction of the dependent variable also known as target variable with respect to the independent variable(s) also known as the predictor variables. This type of regression modelling is used for predictions and forecasting. Regression techniques are mostly categorized into 2 techniques based on the number of independent variables and the type of relationship between the independent and dependent variables. These techniques are:

a) Simple Linear Regression

b) Multiple Linear Regression.

a) **Simple Linear Regression (SLR)**: In this type of regression analysis, the number of independent variables is one and it is assumed that there is a linear relationship between the independent variable and dependent variable. Linear regression consists of finding the best-fitting straight line through the given data points. The best-fitting line is called a ***regression line***. From the mathematical point of view, we know the equation of the line is given by:

$$\boldsymbol{y = mX + c} \text{ .... \{eq.1\}}$$

where **y** is the dependent variable/target variable, **X** is the independent variable/predictor variable, **m** is the slope which determines the angle of the line (also denoted as β) and **c** is the intercept, which determines the value of y when X is 0. Linear regression is just a manifestation of this simple linear equation. Linear regression models are not perfect. They try to approximate the relationship between dependent and independent variables in a straight line and approximation leads to errors, called as ***residuals***. Residual is the error induced in the model and is given by difference between the actual values and the predicted values of y i.e: target variable. It is denoted by *e and* given by,

$$\boldsymbol{e = y_i - y_{pred}} \text{ .... \{eq. 2\}}$$

Here y_i is the actual value and y_{pred} is the predicted value of y. With this assumption eq.1 can now we written as,

$$\boldsymbol{y = \beta_0 + \beta_1 X_i + e_i} \text{ ....\{eq.3\}}$$

where β0 and β1 are two unknown constants that represent the intercept and slope and **e** is the error term. The motive of the linear regression algorithm is to find the best values for β0 and β1 (cost function helps us determine these values). Since we want the best values for β0 and β1, we convert this equation into a minimization equation where we would like to minimize the error between the predicted value and the actual value. Cost function (J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (y_{pred}) and true y value (y_i), and is given by:

$$J = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left[\mathrm{y}_{\mathrm{pred}} - \mathrm{y}_{\mathrm{i}}\right]^2$$

The goal of the linear regression is to find the best-fit straight line from among the given data points and this can be accomplished when the residuals is zero. This is also one of the most important assumptions of Linear Regression that the residuals have mean value of zero. If the expectation (mean) of residuals, e(i), is zero, the expectations of the target variable and the model become the same, which is one of the targets of the model. 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. This is also called as Ordinary Least Square Method and is given by

$$\mathrm{RSS} = \mathrm{e}_1^2 + \mathrm{e}_2^2 + \cdots + \mathrm{e}_\mathrm{n}^2$$

On substituting the values in eq.2 and eq.3 we get:

$$RSS = \sum_{i=1}^{n}(yi - \beta 0 - \beta 1Xi)^2$$

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

i. R^2 or Coefficient of Determination

ii. Residual Standard Error (RSE)

i. R^2 or Coefficient of Determination : R^2 is a measure 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.Higher the R-squared, the better the model fits your data. Mathematically, it is represented as:

$$\boldsymbol{R^2} = 1 - \left(\frac{RSS}{TSS}\right)$$

Where RSS is the Residual Sum of Squares and TSS is Total Sum of Squares (sum of errors of the data points from mean of response variable)

ii. **Residual Standard Error (RSE):** RSE is a measure of lack of fit of the model to the data at hand. If the RSE value is very close to the actual outcome value, then your model fits the data well. If there is a large difference between the values, then the model does not fit the data well.

2) Multiple Linear Regression: In this type of regression analysis, more than one independent variables are used to predict the target variable and it is assumed that there is a linear relationship between the independent variables and dependent variable. It is denoted mathematically as,

$$yi = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + e$$

Where yi is the dependent variable, *X*i are predictor variables, *β*0 is y intercept (constant term), *βn* is slope coefficients for each predictor variable, e is the error term or residual. All the assumptions for simple linear regression hold true for multiple linear regression too. The only thing to look into in terms of Multiple Linear Regression is the multicollineairty between the variables.

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

Ans: Following are the assumptions of linear regression:

a) **Normality assumption for the residuals**: It is assumed that the error terms, e(i), are normally distributed. If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.

b) **Zero mean assumption for residuals**: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero. If the mean of residuals, e(i), is zero, the expectations of the target variable and the model become the same, which is one of the targets of the model.

c) **Constant variance assumption for residuals**: It is assumed that the residual terms have the same but unknown variance, σ2. This assumption is also known as the assumption of homogeneity or homoscedasticity.

d) **Independent error assumption for residuals**: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it implies that there is some relation that the regression model is not able to identify.

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

Ans: **Correlation coefficient**, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson correlation coefficient. It is denoted by r and is mathematically given by,

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

where n is the number of pairs of data. The value of Correlation coefficient (r) lies between -1 and +1. The positive sign indicates a positive relation between the variables i.e: if one variable increases the other variable also increases. The negative sign indicates a negative relation between the two variables where if one variable increases the other decreases. If r is close to +1 or -1 then it indicates strong relation between the variables and depending upon the sign it indicates how the variables are affected. If this value is 0 then it means there is a random, nonlinear relationship between the variables. A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak.

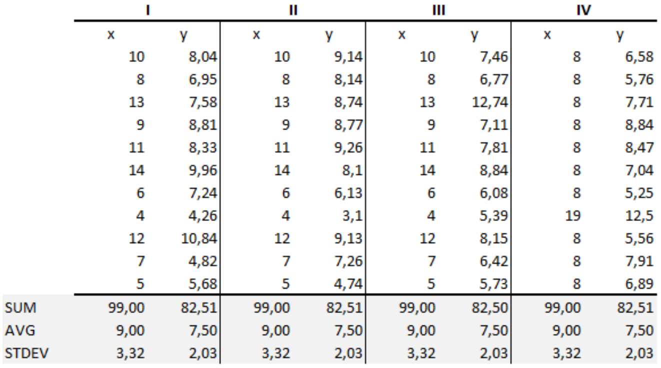
Coefficient of Determination measures the proportion of the variance in the dependent variable that is predicted from the independent variable(s). It is a measure that allows us to determine how certain one can be in making predictions from a certain model. The coefficient of determination is the ratio of the explained variation to the total variation. It is denoted by R^2 and is mathematically represented as:

$$R^2 = 1 - \left(\frac{RSS}{TSS}\right)$$

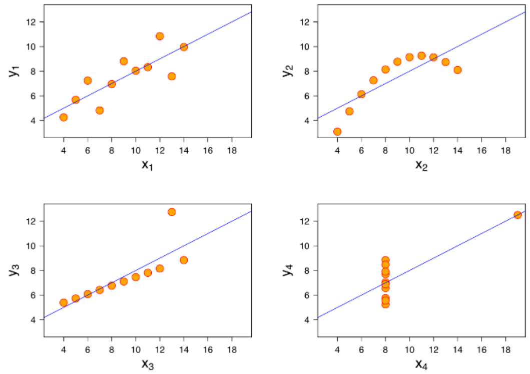
where RSS is the Residual Sum of Square and TSS is Total Sum of Square. The values of R^2 lies between 0 and 1. The coefficient of determination represents the percent of the data that is the closest to the line of best fit. e.g: If the value of R^2 is 0.86 then it means that the model is able to explain 86% of variance in the data i.e: 86% of the total variation in dependent variable can be explained by the linear relationship between independent and the dependent variable.

4. Explain the Anscombe's quartet in detail.

Ans: During modelling, the summary statistics allow us to describe a dataset using just a few key values. This eases out on how and which variables can be considered to obtain a good model. But sometimes relying on Summary statistics can be dangerous since it can be misleading. Even if the summary statistics may seem similar, the overall data distribution can be different. This can be more properly visualized than just checking some numbers. This irregularity in the data pattern is explained by Anscombe's quartet. This quartet emphasizes the importance of data visualization in Data Analysis.



	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 Anscombe's quartet consists of four datasets, each containing 11 (x ,y) pairs. The important thing about these datasets is that they share the same descriptive statistics. As seen from the adjacent dataset, the mean of all the four datasets for x values is 9,00 and for y values is 7,50. Also the standard deviation for these datasets for x is 3,32 and y is 2,03. All the four datasets seem to have the same descriptive Statistic summary. But the story changes when we plot this data. As seen in the adjacent plot, plot 1 appears to have clean and well-fitting linear model. Plot 2 is not distributed normally. The function seems more to be of quadratic type then linear. So if we use linear algorithms to train the data, predictions will be off by huge error. In Plot 3 the distribution is almost linear, but the calculated regression line is thrown off by an outlier. In this case, if we filter out the outlier, the regression line will fit the data more accurately. In plot 4, the data seems mostly constant and is thrown off by just one outlier. So even though the descriptive statics are same for the dataset, the plots show complete different story and based on that we can make changes to our model to increase its accuracy. This is why Data Visualization is one of the most important step in Data Analysis and Ascombe's Quartet just seconds this requirement.

5. **What is Pearson's R?**

Ans: Pearson's correlation coefficient is a measure of the strength of the linear association between the two variables. It is denoted by 'r'. It is a measure of wellness for the data points and how accurately they fit model/line of best fit. It can take values of range +1 to -1. A value greater than 0 indicates a positive correlation i.e. if value of one variable increases the value of other variable are also increases. A value less than 0 indicates a negative correlation i.e. if value of one variable increases the other value of other variable decreases. The value 0 indicates that there is no linear correlation between the the two variables.

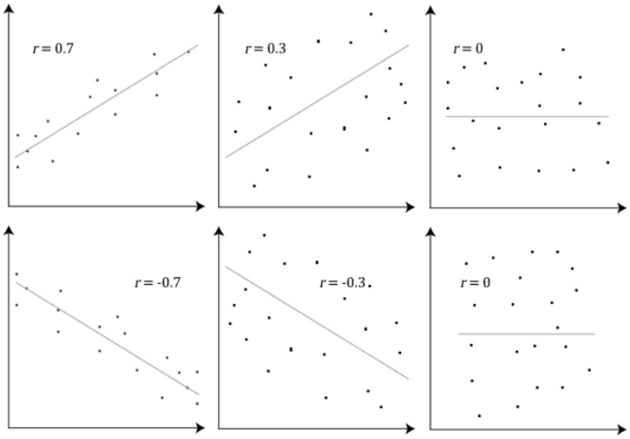
Mathematically it is denoted by.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Where n denotes the number of pairs

∑xy indicates number of product of paired scores, ∑x indicates sum of x scores, ∑y indicates sum of y scores, $\sum x^2$ indicates sum of sqaure x scores and ∑y indicates sum of sqaure of y score.

If the value of r is +1 or -1 means that all the data points are included on the line of best fit i.e. there are no data points that show any variation away from this line. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the adjacent diagram. The value of r=0.7 shows that data points are close to model line and has positive relation similarly r = -0.3 shows negative relation and the data points do not fit the model line properly.



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

Ans: For data analysis, in a given dataset, the range of values may vary widely in magnitude, units and range causing issues during the processing/modelling the data. The results of predictions using this raw data can vary greatly, hence to reduce this error it is beneficial to bring all the data to the same level or range. The method use to standardize the range of features is called *Feature Scaling* or just *Scaling*.

Consider an example of a dataset having entries for weight of particular product. The dataset has few entries in terms of grams while some in term of kilograms. The offset between these values is approximately 1000. So suppose one entry has value 5kg and other has 5000gms even though both are similar in terms of weight they will be treated differently during calculation since both these individual entries have different values inducing the unnecessary error. One way to encounter this to convert all the data on the same scale. For this example, either convert all data into gm or kgs. However, what suppose there are multiple metric systems representing the data. It will be very daunting to analyze the data for large datasets. To reduce such type of errors scaling is used while pre – processing the data. In scaling, the data is transformed in such a way that the features are within a specific range i.e. 0 and 1. It is mathematically denoted as

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

where x' is the normalized value, x is the actual values, x_{min} is the minimum value of x and x_{max} is the max value of x. There are 2 standard methods for scaling the data: **Normalization and Standardization**.

In **normalization**, the values are modified in such a way that they can be more described like a normal distribution. Normal distribution or Gaussian distribution, which is also known as the **bell curve**, is a specific statistical distribution where a approximately equal observations from dataset fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean. It is mathematically denoted as:

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

Where x is the value of x from dataset at given point, x_{mean} is the mean of x values and x_{max} $_{and}$ x_{min} are maximum and minimum values of x respectively. For normalization, the maximum value you can get after applying the formula is 1, and the minimum value is 0. So all the values will be between 0 and 1. While using the regression techniques like Linear Regression we will need to normalize the data.

On the other hand, Standardization *also known as z-score normalization* transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1. It is mathematically denoted by:

$$x' = \frac{x - x_{mean}}{\sigma}$$

Here x' is the standardized value, x is the value of x in dataset, x_{mean} is the mean of x values and σ is the standard deviation of x. It is widely used in SVM, logistics regression and neural networks. Standardization can be use in Logistic Regressions.

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

Ans: VIF is also called a Variance Inflation Factor. It detects multicollinearity in regression analysis. Multicollinearity is the correlation between independent variables/predictors in a model. Multicollinearity can adversely affect the regression result. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIF is mathematically represented as:

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

Where R^2 is the R-sqaured value of given predictor.

When the value of R^2 is high , the value of VIF is high too. E.g: If value of R^2 is 0.9 the value of VIF will be 10. 10 is definitely high value. A high value of VIF represents the multicollinearity among the predictor variables. This means that for any given variables x1, x2, x3, x4 and x5, suppose the value of VIF is high for x1, then x1 is correlated with other predictor variables like x2 and x3 or x4 and x5 and so on. i.e: the variable with high VIF is correlated with other variable combination and hence can be eliminated from the analysis. The value of VIF will be infinite if value of R^2 is 1. If R^2 is 1 that means that the variable into consideration is completely correlated with some other variable in model and hence is insignificant for analysis and can be removed.

8) What is the Gauss-Markov theorem?

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

Following are the assumptions which if true then Guass Markov Theorem gives the best linear unbaised estimate for regression model:

1) Model should be linear: It is assumed that there should be a linear relationship between the dependent and independent variables. The model should be represented in the form of:

$$\mathbf{y = \beta_0 + \beta_1 X_i + \mathit{e}_i}$$

where β0 and β1 are two unknown constants that represent the intercept and slope and **e** is the error term.

2) Residual should be normally distributed: It is assumed that the error terms/residuals are normally distributed.

3) The error term has a population mean of zero: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

4) Homoscedaticity: It is assumed that the residual terms are homogeneous and have same but unknown variance. (Homoscedasticity)

5) Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

6) Assumptions about the estimators: The independent variables are linearly independent of each other and are measured without error.

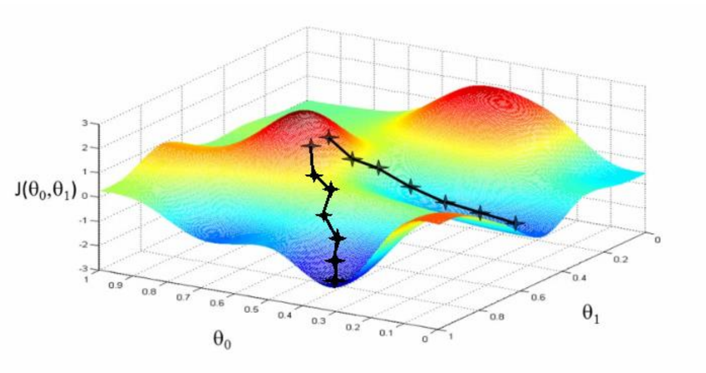
If all these six assumptions are met then we can say that the Ordinary Least Squares estimates for the regression coefficients will give the best linear unbaised estimate.

9) Explain the gradient descent algorithm in detail

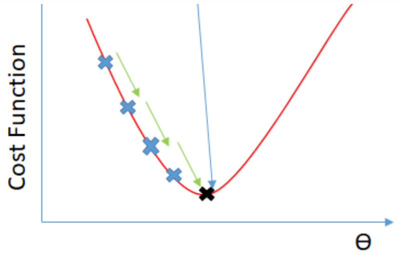
Ans: Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. Optimization means getting the optimal output for a particular function. In mathematical term optimization is the task of minimizing/maximizing an objective function f(x) parameterized by x.

Gradient Descent is the most common optimization algorithm in machine learning. It is a first-order optimization algorithm which takes into account the first derivative while performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function with respect to the parameters where the gradient gives the direction of the steepest ascent (since we need to find the steepest descent or minimum point).

To represent this graphically consider the adjacent cost space plot. Cost space plot is the plot of how the algorithm will behave when we choose a particular value for the parameter. The plot has peaks and valleys. The ultimate goal is to reach from peak (high cost) to the bottom of the valley (lost cost) from any given point in the space following the directions as shown. (Goal is to reach the local minimum)



Here J(ϴ0, ϴ1) is the cost function and ϴ0 and ϴ1 are the variables which the cost function will determine. We start with calculating the gradient at the given point and move in the direction of the negative gradient. Once we move to the next point we recalculate the gradient at this new point and move in the direction on negative gradient. We do this until the algorithm reaches to the minimum point and the algorithm converges.

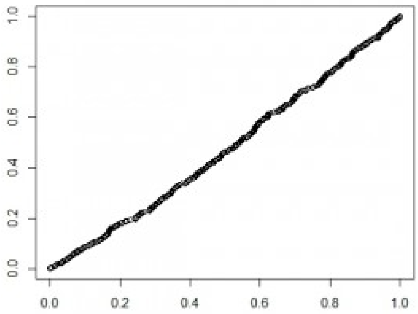


The size of the step we take on each iteration to reach the local minimum is determined by the learning rate α. We follow the direction of the slope descent until we reach a local minimum.Determining the learning rate, α, is crucial step in the gradient descent algorithm because if α is very small it would take long time to reach the local minimum or for algorithm to converge and will become computationally very much expensive. While if α is large it may overshoot the minimum and fail to converge at the right spot. Therefore selecting a learning rate is extremely important. It can be accomplished by plotting the cost function J(w) against different values of α and picking up good value of α before its first converge. The most common used learning rates are: 0.001, 0.003, 0.01, 0.03, 0.1 and 0.3

Another important thing to consider before deciding the learning rate is that the data should be normalized. This is important because if the data is on different scales, the width of contours (levels of constrast) will be long and narrow and the algorithm will take huge time to converge. The data should be normalized such that mean is 0 and standard deviation is 1. This can be accomplished by formula $(x_i - \mu) / \sigma$, where x_i is the value of x at given instance, μ is mean of x and σ is standard deviation of x.

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

Ans: Q-Q (Quantile to Quantile) plot is the plot of two quantiles against each other. A quantile is a fraction where certain values fall below a particular range. For example, a median is a quantile where 50% of the data lie in the lower part while other 50% lies in the upper part. The best way to view the quantiles is the box plots where there are 3 quantiles shown (25%, 50% and 75%). Quantile plots are used to check if the given two datasets come from the same distribution. A 45 degree angle straight line can be seen if the two datasets come from the same distribution somewhat as shown in the adjacent figure.



In linear regression, assumption of normality is one of the assumptions made during modelling. Assumption of normality states that in the residuals/error terms are normally distributed i.e. we are sampling from a normally distributed population. We can check either this by plotting a histogram and checking if all the values lie between the bell curve or we can check this by plotting the Q-Q plots. So Q-Q plots can be used in linear regression to verify the error distribution and thereby verifying assumption of normality for the linear modelling.