upGrad

International Institute of Information Technology Bangalore

Linear Regression Assignment – Subjective Questions

Author: Anish Mahapatra
E-mail id: anishmahapatra01@gmail.com

Q1: Explain the Linear Regression algorithm in detail.

A. Linear regression is a method of finding the best straight-line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Residual Sum of Squares Method.

Hypothesis function for Linear Regression is as follows:

$$\text{Hypothesis:} \qquad h_\theta(x) = \theta_0 + \theta_1 x$$

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

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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 θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$\text{Cost Function:} \quad J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

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

To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

Q2: What are the assumptions of linear regression regarding residuals?

A. The assumptions of linear regression regarding residuals are as follows:

1. **Normality assumption:** It is assumed that the error terms, $\varepsilon^{(i)}$, 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

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. The mean of the residuals should be zero for the following equation:

   $$Y^{(i)i} = \beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)}$$

   This is the assumed linear model, where ε is the residual term.

   $$E(Y) = E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$$
   $$= E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$$

   If the expectation(mean) of residuals, $E(\varepsilon^{(i)})$, is zero, the expectations of the target variable and the model become the same, which is one of the targets of the model.

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. A random variable is said to be heteroscedastic when different subpopulations have different variabilities (standard deviation). The existence of heteroscedasticity gives rise to certain problems in the regression analysis as the assumption says that error terms are uncorrelated and, hence, the variance is constant. The presence of heteroscedasticity can often be seen in the form of a cone-like scatter plot for residual vs fitted values.

4. **Independent error assumption:** It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero. The residuals (also known as the error terms) should be independent, meaning there is no correlation between the residuals and the predicted values, or among the residuals. Any correlation implies that there is some relation that the regression model is not able to identify.

Q3. What is the coefficient of correlation and the coefficient of determination?

A. Linear correlation coefficient measures the strength and the direction of a linear relationship between two variables. The formula for correlation coefficient is as follows:

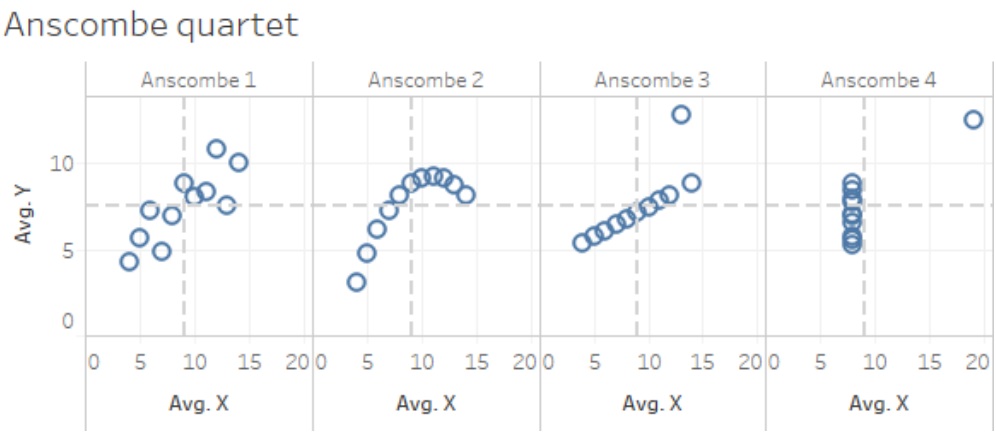
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

The value of r is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively. If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase. If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease. If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables. A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*.

The *coefficient of determination,* r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The *coefficient of determination* is the ratio of the explained variation to the total variation. It is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y. The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y. The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y.

Q4: Explain the Anscombe's quartet in detail.

A. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x, y) fixed points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.
The visualization and explanation of each segment of the quartet is as follows:



Anscombe 1 – This graph shows a simple linear positive relationship. It is what we would expect to see, assuming a normal distribution.

Anscombe 2 – This graph does not appear to be normally distributed. We can however see a relationship between the 2 variables, it appears to be quadratic or parabolic, but it is not linear.

Anscombe 3 – This graph is showing a clear outlier in the dataset. The data points, except for the outlier are showing what appears to be perfect linear relationship, but because of the outlier the value of the correlation coefficient has been reduced from 1 to 0.816.

Anscombe 4 – In this graph we can see that the value of x stays constant except for one outlier. This outlier has created the same correlation coefficient as the other datasets, which is a high correlation, however the relationship between the two variables is not linear.

Significance of **Anscombe's quartet** are as follows:

1. Data visualization is crucial in developing a sensible statistical model
2. Pearson correlation coefficient (PCC) only captures linear relationships. Therefore, it is irrelevant in non-linear cases
3. PCC and linear regression are extremely sensitive to outliers

Q5: What is Pearson's R?

A. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
This is like the answer described in question 3.

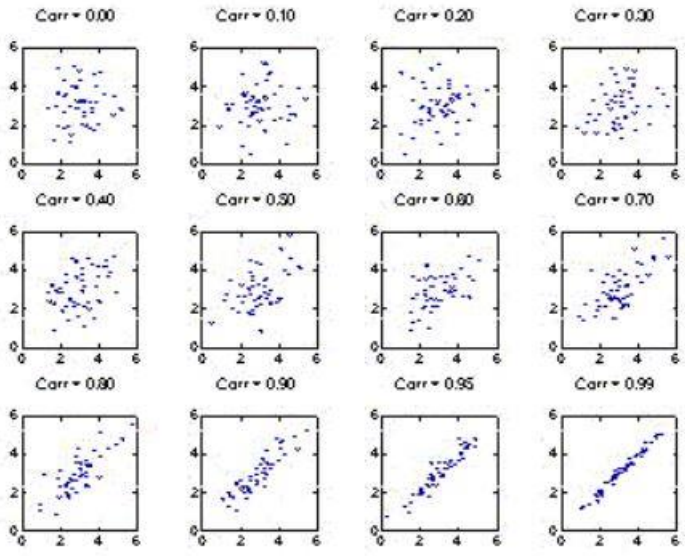
The significance of different values of Pearson's coefficient is as follows:

r = -1: Data lies on a perfect straight line with a negative slope

r = 0: No linear relationship between the variables

r= +1: Data lies on a perfect straight line with a positive slope

The different visualizations of linear relationship based on r value is shown below:



Q6: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A. In scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].

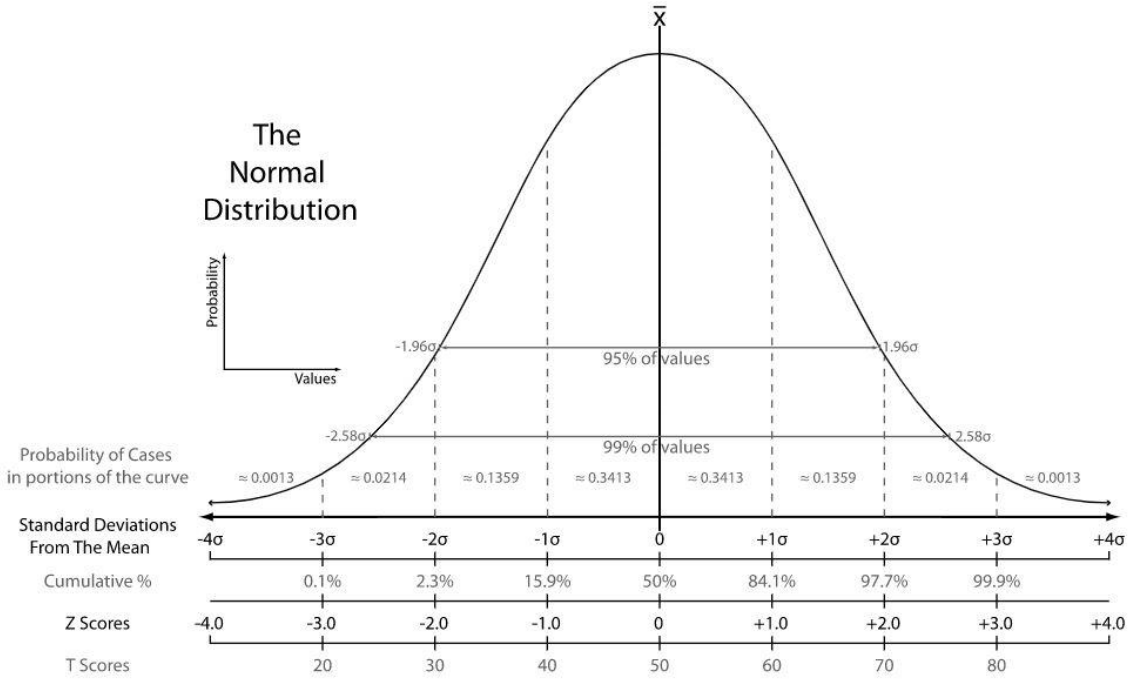
$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Here x' is the normalized value

Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbours (KNN) where distance between the data points is important. For example, in the dataset containing prices of products; without scaling, SVM might treat 1 USD equivalent to 1 INR though 1 USD = 65 INR.

Difference between normalized scaling and standardized scaling:

The point of normalization is to change your observations so that they can be described as a normal distribution. Normal distribution (Gaussian distribution), also known as the **bell curve**, is a specific statistical distribution where a roughly equal observations fall above and below the mean, the mean and the median are the same, and there are more observations closer to the mean. It looks as follows:



Normalization, it's just another way of normalizing data. Note that, it's a different from min-max scaling in numerator, and from z-score normalization in the denominator.

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

For normalization, the maximum value you can get after applying the formula is 1, and the minimum value is 0. All the values will be between 0 and 1.

Standardization *(also called **z-score normalization**)* transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

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

x is the original feature vector, X_mean is the mean of that feature vector, and σ is its standard deviation.

The z-score comes from statistics, defined as

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

μ is the mean. By subtracting the mean from the distribution, we're essentially shifting it towards left or right by amount equal to mean i.e. if we have a distribution of mean 100, and we subtract mean 100 from every value, then we shift the distribution left by 100 without changing its shape. Thus, the new mean will be 0. When we divide by standard deviation σ, we're changing the shape of distribution. The new standard deviation of this standardized distribution is 1 which you can get putting the new mean, μ=0μ=0 in the z-score equation.

It's widely used in SVM, logistics regression and neural networks.

Q7: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

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

If all the independent variables are orthogonal to each other, then VIF = 1.0. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Q8: What is the Gauss-Markov theorem?

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

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

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

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

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

$y_i = x_i'\beta + \varepsilon_i$

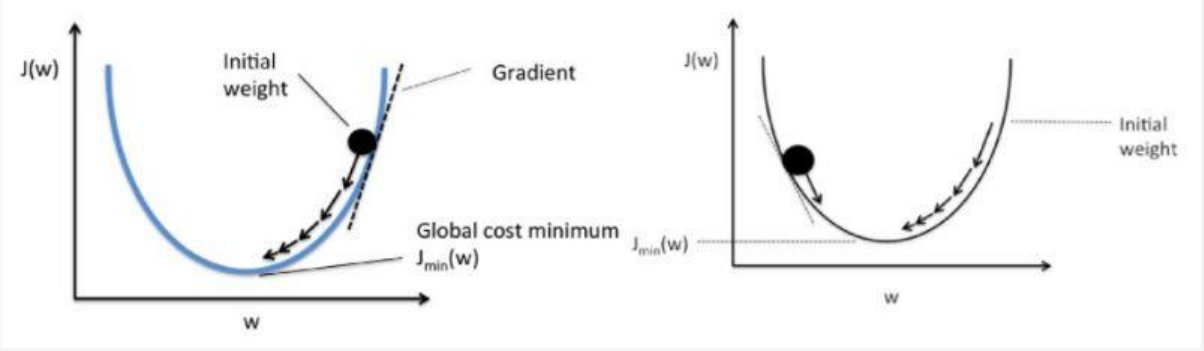
and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

- $E\{\varepsilon i\} = 0, i = 1, \ldots, N$
- $\{\varepsilon 1 \ldots\ldots \varepsilon n\}$ and $\{x1\ldots..,xN\}$ are independent
- $cov\{\varepsilon i, \varepsilon j\} = 0, i, j = 1,\ldots., N\ I \neq j.$
- $V\{\varepsilon 1 = \sigma 2, i= 1, \ldots.N$

Q9: Explain the gradient descent algorithm in detail

A. Gradient descent is an optimisation algorithm. In linear regression, it is used to optimise the cost function and find the values of the βs (estimators) corresponding to the optimised 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).



Mathematically, the aim of gradient descent for linear regression is to find the solution of

ArgMin $J(\Theta_0,\Theta_1)$, where $J(\Theta_0,\Theta_1)$ is the cost function of the linear regression. It is given by:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Here, *h* is the linear hypothesis model, $h = \Theta_0 + \Theta_1 x$, *y* is the true output, and *m* is the number of datapoints in the training set.

Gradient descent starts with a random solution, and then, based on the direction of the gradient, the solution is updated to the new value, where the cost function has a lower value.

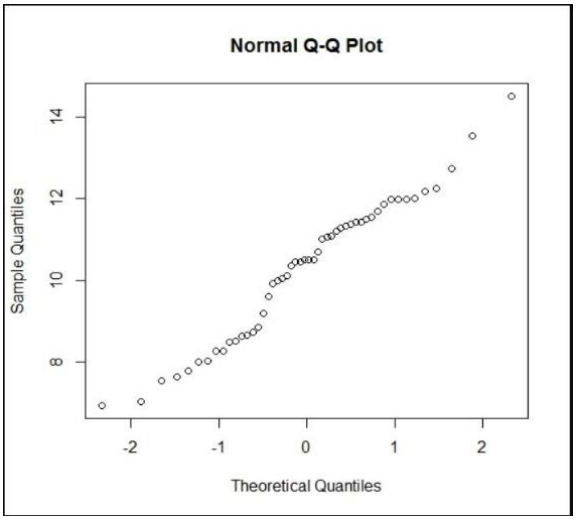
The update is:

Repeat until convergence:

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}).\, x_j^{(i)} \quad \text{for j = 1,2,...,n}$$

Q10: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A. 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 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. Following is an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Q-Q plots can be used to check the homoscedasticity and normality assumptions of **linear regression**.

A residual plot lets you see if your data appears homoscedastic. Homoscedasticity means that the residuals, the difference between the observed value and the predicted value, are equal across all values of your predictor variable. If your data are homoscedastic then you will see the points randomly scattered around the x axis. If they are not (e.g. if they form a curve, bowtie, fan etc.) then it suggests that your data doesn't meet the assumption.

Q-Q plots let you check that the data meet the assumption of normality. They compare the distribution of your data to a normal distribution by plotting the quartiles of your data against the quartiles of a normal distribution. If your data are normally distributed, then they should form an approximately straight line.