Assignment – Linear Regression
Subjective Questions & Answers

1. Explain the linear regression algorithm in detail.

Answer:

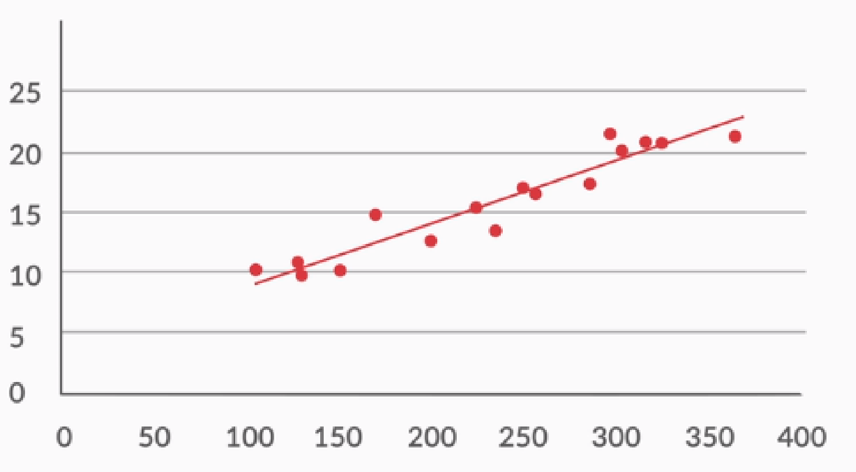
Linear regression machine learning algorithm is a supervised machine learning algorithm model, which is used to determine the strength of relationship between a scalar (dependent) variable and an explanatory (independent) variable.

> *By definition, it is a technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with, and depends on, a change in one or more independent variables.*

> *The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon where heights of descendants of tall ancestors tend to regress down towards a normal average (mean).*

More importantly, linear regression model works based on a dependent variable and an independent variable concept where the values of the dependent variable is predicted based on the independent variable by finding the linear relationship between them.

To find the linear relationship between dependent and independent values, dependent variable is placed on Y-axis and independent variable is placed on the X-axis. If we imagine a scatter plot with all the data point between a dependent variable on X-axis and independent variable on Y-axis, a straight line drawn through these points to predict a new value of dependent variable.



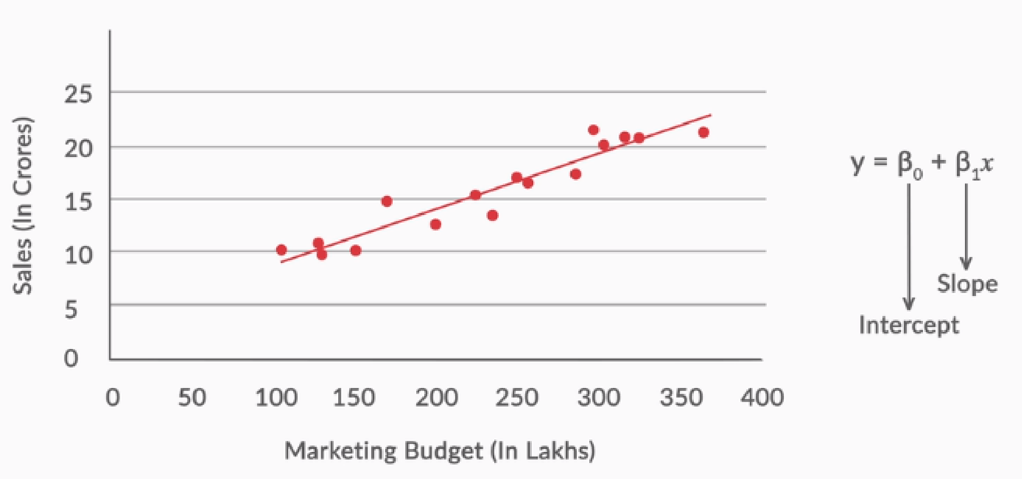
We need to find a best fit line through these data points to predict accurate values; and to determine that we use the least squares method which is a statistical technique to determine the line of best fit for a model, specified by an equation with certain parameters to observed data. The least squares method provides the foundation, for the placement of best-fit line among the data points being studied.

Assignment – Linear Regression
Subjective Questions & Answers

Here come the importance to know the equation of a straight line which is y=mx+c, where y is the dependent variable, m is slope(y2-y1/x2-x1) and c is the constant where the straight line touches the x axis.

Therefore the equation of the best fit regression line is given below. In statistical terms this equation is translated as β's.

$Y = \beta_0 + \beta_1 X$ Where y is the dependent variable, β1 is slope (y2-y1/x2-x1) and β0 is the constant where the straight line touches the Y axis.



Therefore the formula for linear regression is given as below:

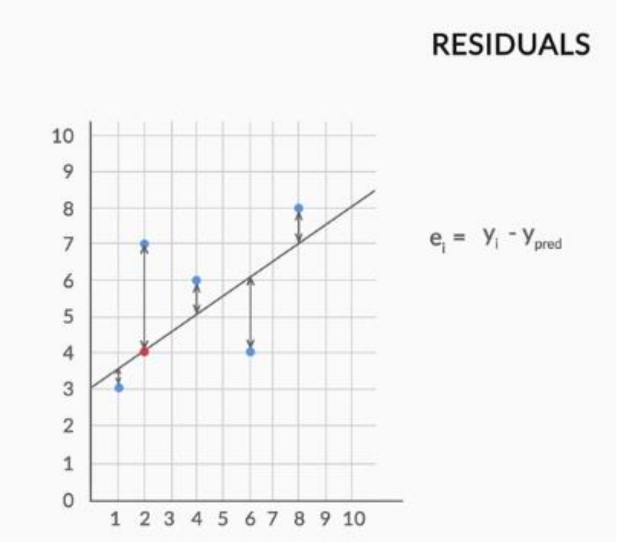
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,\ i = 1, \ldots, n.$$

Residual Sum of Squares (RSS):

The best-fit line is achieved by minimizing the residual sum of squares (RSS).

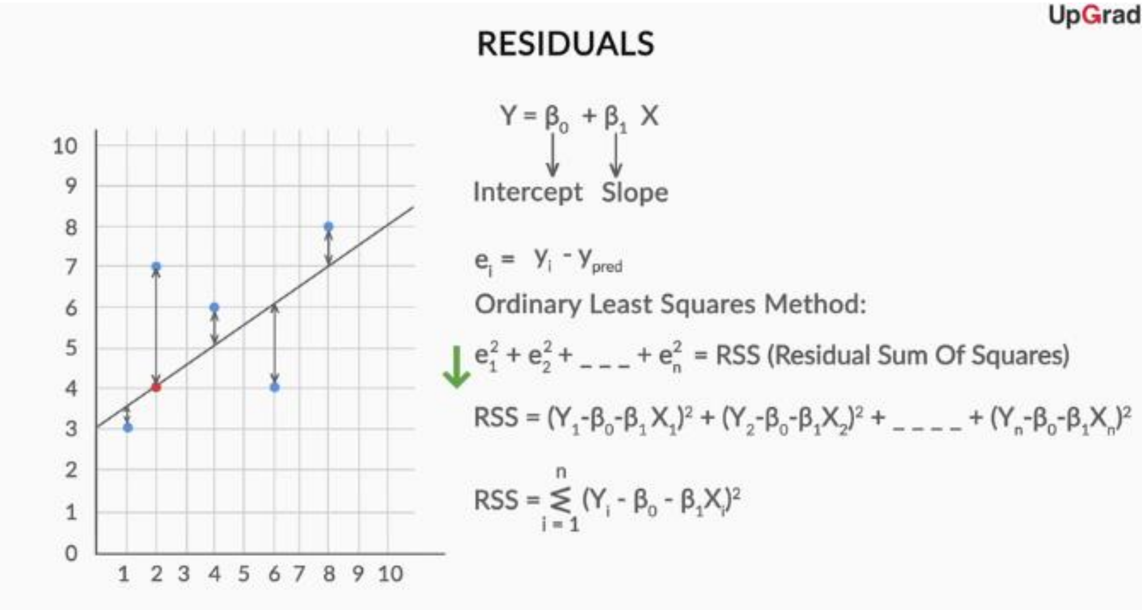
The most common application of this method, which is sometimes referred to as "linear" or "ordinary", aims to create a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model.

Assignment – Linear Regression
Subjective Questions & Answers


RESIDUALS

$e_i = y_i - y_{pred}$

The line of best fit determined from the least squares method has an equation that tells the story of the relationship between the data points. Line of best fit equations may be determined by computer software models, which include a summary of outputs for analysis, where the coefficients and summary outputs explain the dependence of the variables being tested.


UpGrad

RESIDUALS

$Y = \beta_0 + \beta_1 X$

Intercept Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_2-\beta_0-\beta_1 X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1 X_n)^2$

$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

Application of linear regression model:

Linear regression model is applicable in areas such as

1. Prediction,
2. Forecasting,
3. Process optimization
4. Extract new insights from data.

Assignment – Linear Regression
Subjective Questions & Answers

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

Answer:

Assumption on residuals:

Independence of errors (error term is additive -No Interactions):
If there were dependence in errors, which means errors are capturing some information about the model. This will in turn leads to Inaccurate model.

Residuals should be normally distributed:
If not, our model is not consistent across the full range of your observed data (This assumption may be checked by looking at a **Histogram or a Q-Q-Plot**.)
Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves.

The mean of residuals should be zero:
If residuals mean is not zero which it carry some information about the dependent variable.

Variance of error terms should be similar across the values of the independent variables(Homoscedastic)
A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.
If the data are **heteroscedastic**, a non-linear data transformation or addition of a quadratic term might fix the problem.

No Auto Correlation among residuals:
If there is auto correlation among residuals which means current value is dependent on the historic values and that there is a definite unexplained pattern in the Independent variable that shows up in the disturbances.It can be validated using Durban watson test

Assignment – Linear Regression
Subjective Questions & Answers

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

Answer:

Coefficient of Correlation:

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; • 0 means not at all correlated.
Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable
For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That's why R square is a better term. You can explain R square for both simple linear regressions and also for multiple linear regressions

Coefficient of determinations:

Coefficient of determinations used to explain how much variability of one factor can be caused by its relationship to another factor and is represented as a value between 0 and 1. This correlation is known as the **"goodness of fit."**.
The **"goodness of fit"**, or the degree of linear correlation, measures the distance between a fitted line on a graph and all the data points that are scattered around the graph.
A value of 1.0 indicates a perfect fit, and it is thus a very reliable model for future forecasts, indicating that the model explains all of the variations observed.
A value of 0, on the other hand, would indicate that the model fails to accurately model the data at all.
For a model with several variables, such as a multiple regression model, the adjusted R2 is a better coefficient of determination.

Assignment – Linear Regression
Subjective Questions & Answers

4. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. Below are the pointers explaining in detail.

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 can not accurately portray data in its native form.
2. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.
3. 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.

5. What is Pearson's R?

Answer:

The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables.

It is referred to as Pearson's correlation or simply as the correlation coefficient. 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.

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables.

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

Answer:

What is scaling?

Scaling is the assignment of objects to numbers according to a rule. In scaling, the objects are text statements, usually statement of attitude, opinion or feeling. Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

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 Eucledian distance between two data points in their computations, this is a problem.
- If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.
- More importantly, We can speed up gradient descent by scaling. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

What is the difference between normalized scaling and standardized scaling?

Normalization: Normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

Standardization: typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Answer:

An infinite (inf) VIF will be returned for two variables that are exactly collinear, variables that are exactly the same or linear transformations of each other.

8. What is the Gauss-Markov theorem?

Answer:

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

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

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.

Erogeneity: 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.

9.

10. Explain the gradient descent algorithm in detail.

Answer:

Introduction:

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.

Explination:

Below is an example to illustrate gradient descent model:

For instance, you are at the top of a mountain, and you have to reach a lake which is at the lowest point of the mountain (a.k.a valley). A twist is that you are blindfolded and you have zero visibility to see where you are headed. So, what approach will you take to reach the lake?

The best way is to check the ground near you and observe where the land tends to descend. This will give an idea in what direction you should take your first step. If you follow the descending path, it is very likely you would reach the lake.

Suppose we want to find out the best parameters (θ1) and (θ2) for our learning algorithm. Similar to the analogy above, we see we find similar mountains and valleys when we plot our “cost space”. Cost space is nothing but how our algorithm would perform when we choose a particular value for a parameter.So on the y-axis, we have the cost J(θ) against our parameters θ1 and θ2 on x-axis and z-axis respectively.

In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

- On the basis of differentiation techniques
  1. First order Differentiation
  2. Second order Differentiation

Summary:

Gradient descent requires calculation of gradient by differentiation of cost function. We can either use first order differentiation or second order differentiation.

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

Answer:

Q-Q plot is used to assess if your residuals are normally distributed.

Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don’t, your residuals aren’t Gaussian and thus your errors aren’t either.

This implies that for small sample sizes, you can’t assume your estimator beta is Gaussian either, so the standard confidence intervals and significance tests are invalid.