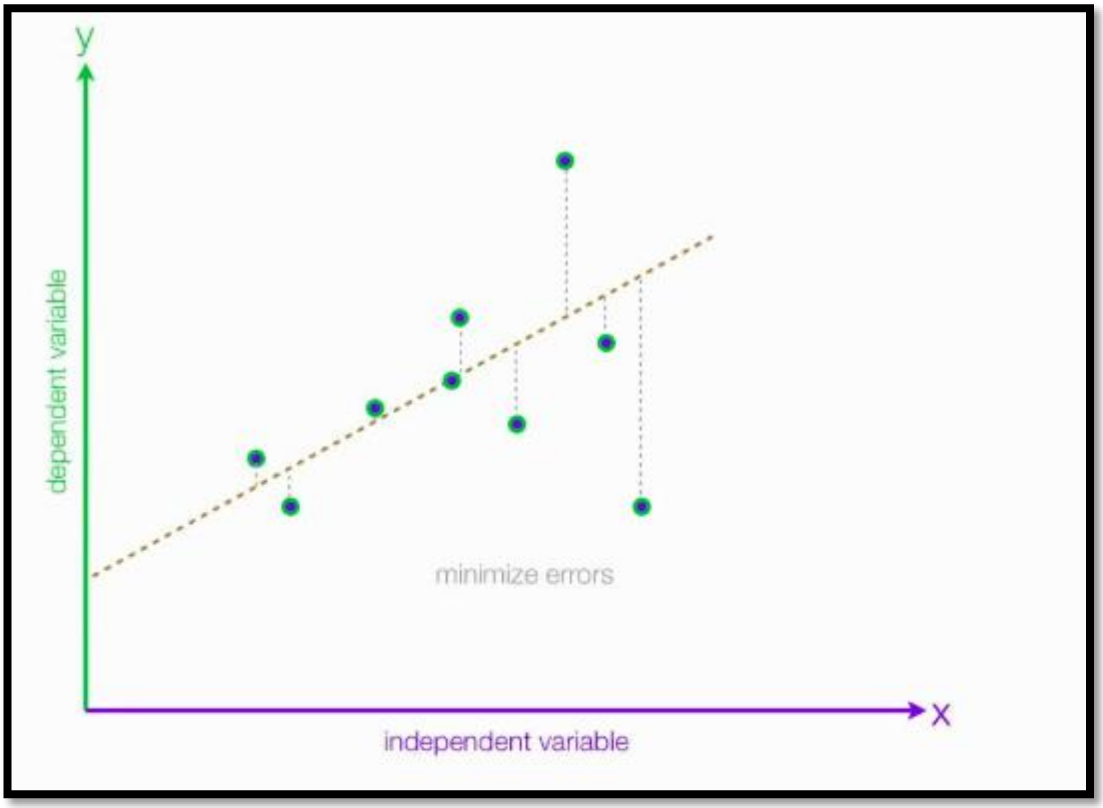
RAVI EDLA – Linear Regression Subjective Questions Answers

1. **Explain the linear regression algorithm in detail.**
   Regression algorithms fall under the family of Supervised Machine Learning algorithms. Supervised learning algorithms models the dependencies and establish relationships between the target output and input features to predict the value for new data. The data is spilt into train data and spilt data in some ratio (say 70%:30%, 80%:20%) Regression algorithm then predicts output values based on input features of train data and test it on test data.



Linear regression is used to perform tasks to predict the value of dependent variable (y) based on a given independent variable (x). By this regression technique we can find out linear relationship between x (input) and y(output). Hence, it called as Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

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

   The assumptions of linear regression are:

   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:

   ✓ Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.

   ✓ 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.

   ✓ Constant variance assumption: 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.

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

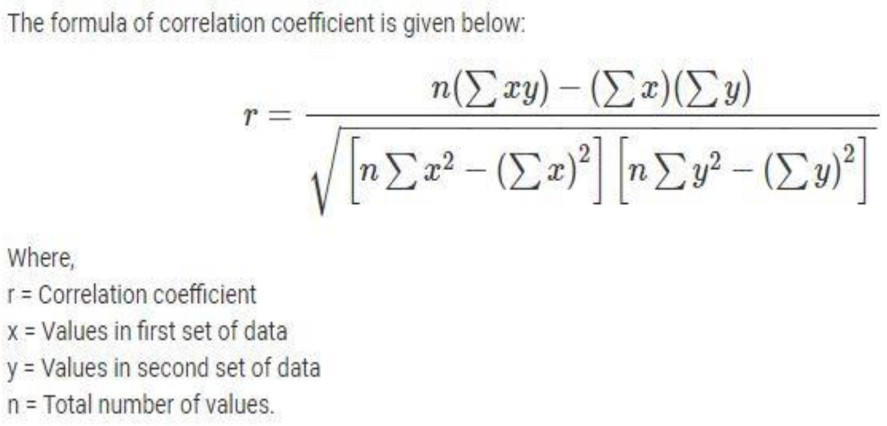
   3) Assumptions about the estimators:

   ✓ The independent variables are measured without error.

   ✓ The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.

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

   The coefficient of determination or R squared method is the proportion of the variance in the dependent variable that is predicted from the independent variable. It indicates the level of variation in the given data set.



The formula of correlation coefficient is given below:

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

Where,
r = Correlation coefficient
x = Values in first set of data
y = Values in second set of data
n = Total number of values.

- The coefficient of determination is the square of the correlation (r), thus it ranges from 0 to 1.
- With linear regression, the correlation of determination is equal to the square of the correlation between the x and y variables.

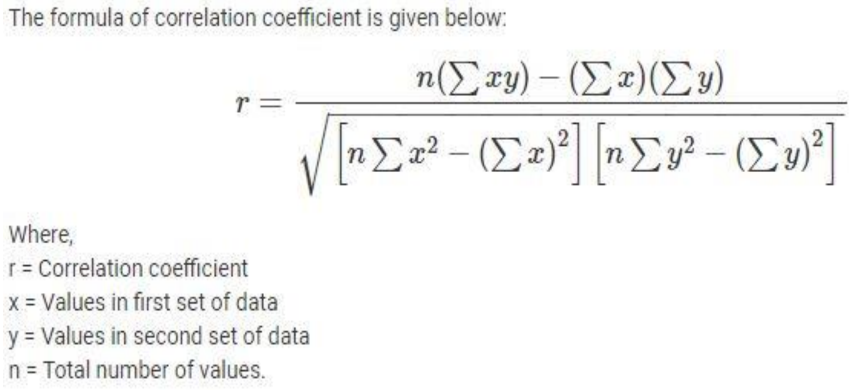
- If R2 is equal to 0, then the dependent variable should not be predicted from the independent variable.
- If R2 is equal to 1, then the dependent variable should be predicted from the independent variable without any error.

If R2 is between 0 and 1, then it indicates the extent that the dependent variable can be predictable. If R2 of 0.10 means, it is 10 per cent of the variance in y variable is predicted from the x variable. If 0.20 means, it is 20 per cent of the variance is y variable is predicted from the x variable, and so on.

The value of R2 shows whether the model would be a good fit for the given data set. On the context of analysis, for any given per cent of the variation, it (good fit) would be different. For instance, in a few fields like rocket science, R2 is expected to be nearer to 100 %. But R2 = 0(minimum theoretical value), which might not be true as R2 is always greater than 0 (by Linear Regression).

The value of R2 increases after adding a new variable predictor. Note that it might not be associated with the result or outcome. The R2 which was adjusted will include the same information as the original one. The number of predictor variables in the model gets penalized. When in a multiple linear regression model, new predictors are added, it would increase R2. Only an increase in R2 which is greater than the expected (chance alone), will increase the adjusted R2

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. It is given by



The formula of correlation coefficient is given below:

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

Where,
r = Correlation coefficient
x = Values in first set of data
y = Values in second set of data
n = Total number of values.

Correlation co-efficient (r) ranges from -1 to 1.

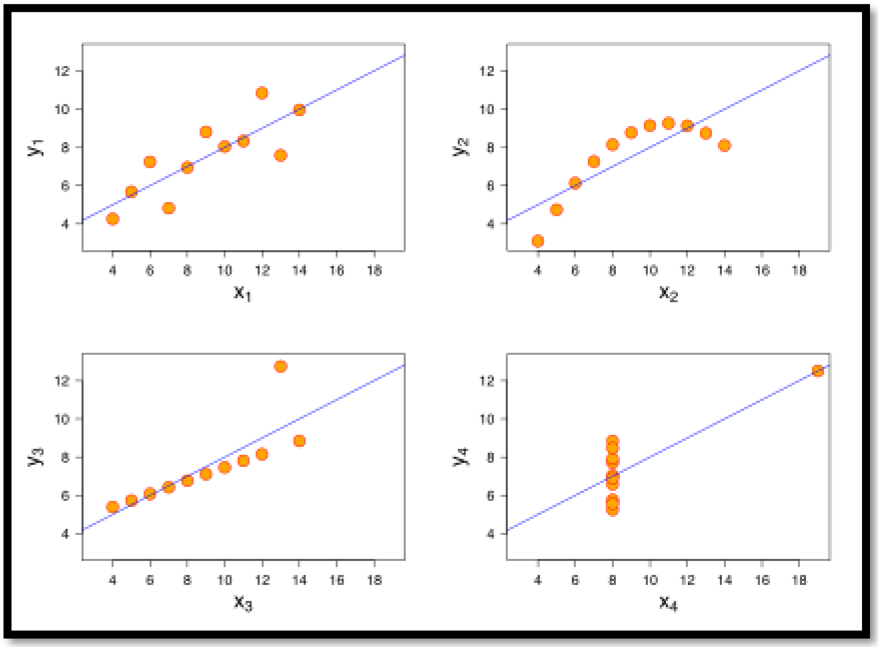
- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length. Also known as **Positive correlation**.

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed. Also known as **Negative correlation**.
- Zero means that for every increase, there isn't a positive or negative increase. It is generally called as **No correlation**. A value near zero means that there is a random, nonlinear relationship between the two variables.

Note that r is a dimensionless quantity; that is, it does not depend on the units employed.
A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line.
If r = +1, then 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. These values can vary based upon the "**type**" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

4. Explain the Anscombe's quartet in detail.

Francis Anscombe 1973 depicted variance in any data set using just four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. These data sets, collectively known as "Anscombe's Quartet," are shown below



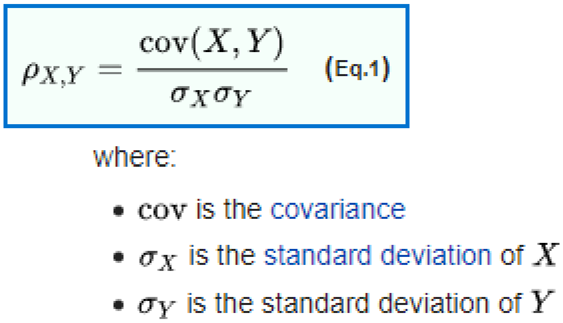
- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables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

The quartet often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

5. What is Pearson’s R?

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.

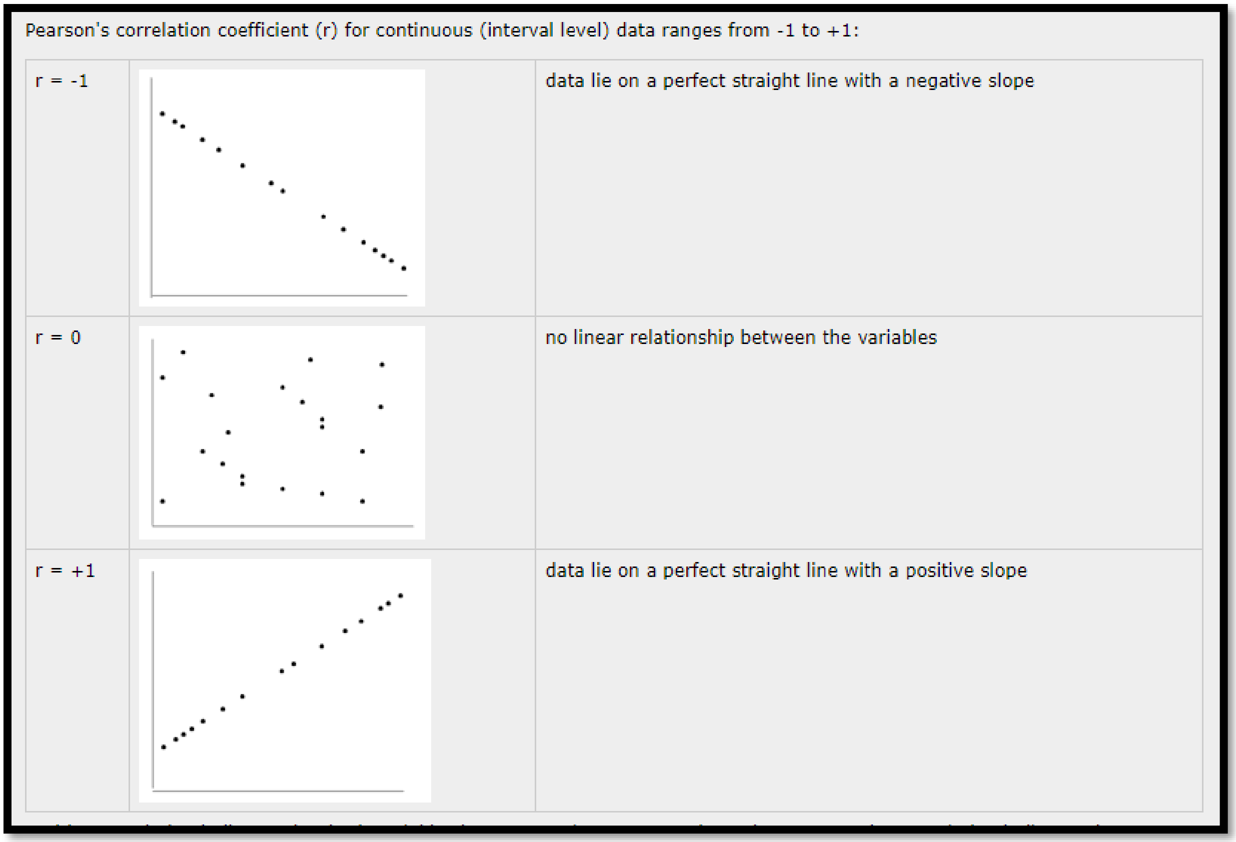
The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. It is defined as



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y



Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

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

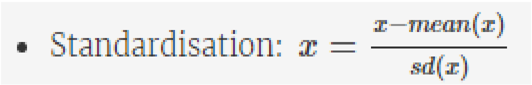
Most of the times, your dataset will contain features highly varying in magnitudes, units and range. The data can vary greatly between different units, example 5kg and 5000gms The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.

To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling

There are four common methods to perform Feature Scaling.

- Standardisation:

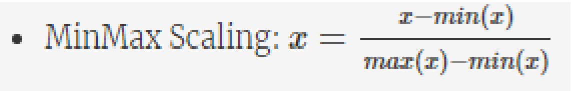
  Standardisation replaces the values by their Z scores.

  - Standardisation: $x = \frac{x - mean(x)}{sd(x)}$

  This redistributes the features with their mean μ = 0 and standard deviation σ =1 . sklearn.preprocessing.scale helps us implementing standardisation in python.

- Min-Max Scaling



- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

This scaling brings the value between 0 and 1.

Difference between normalized scaling and standardized scaling:
Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost. While Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance)

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
If there is perfect correlation, then VIF = infinity.
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.

8. What is the Gauss-Markov theorem?
In statistics, the Gauss–Markov theorem states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero.
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.

Gauss Markov Assumptions:
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

yi = xi‘ β + εi

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

- E{εi} = 0, i = 1, . . . , N
- {ε1……εn} and {x1…..,xN} are independent
- cov{εi, εj} = 0, i, j = 1,…., N I ≠ j.
- V{ε1 = σ2, i= 1, ….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.

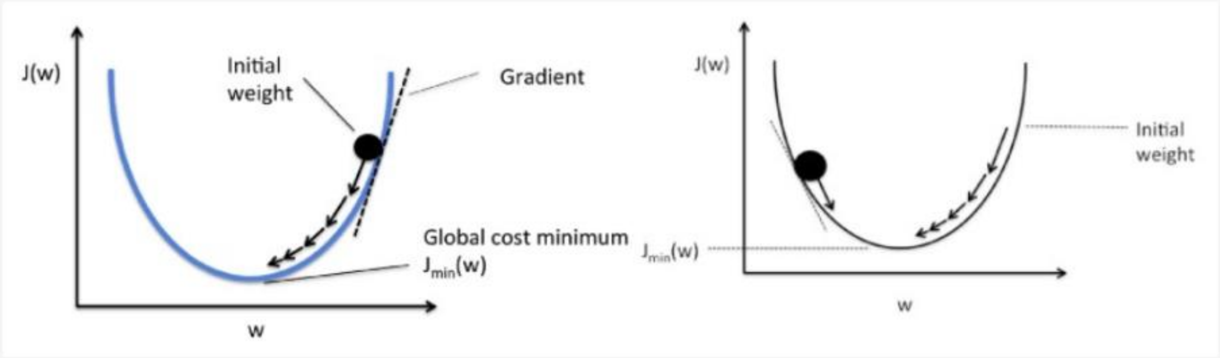
9. Explain the gradient descent algorithm in detail.

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 is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.

In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

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).



Gradient Descent

Steps:

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) = 1N∑i=1n(yi−(mxi+b))2f(m,b)=1N∑i=1n(yi−(mxi+b))2

The gradient can be calculated as:

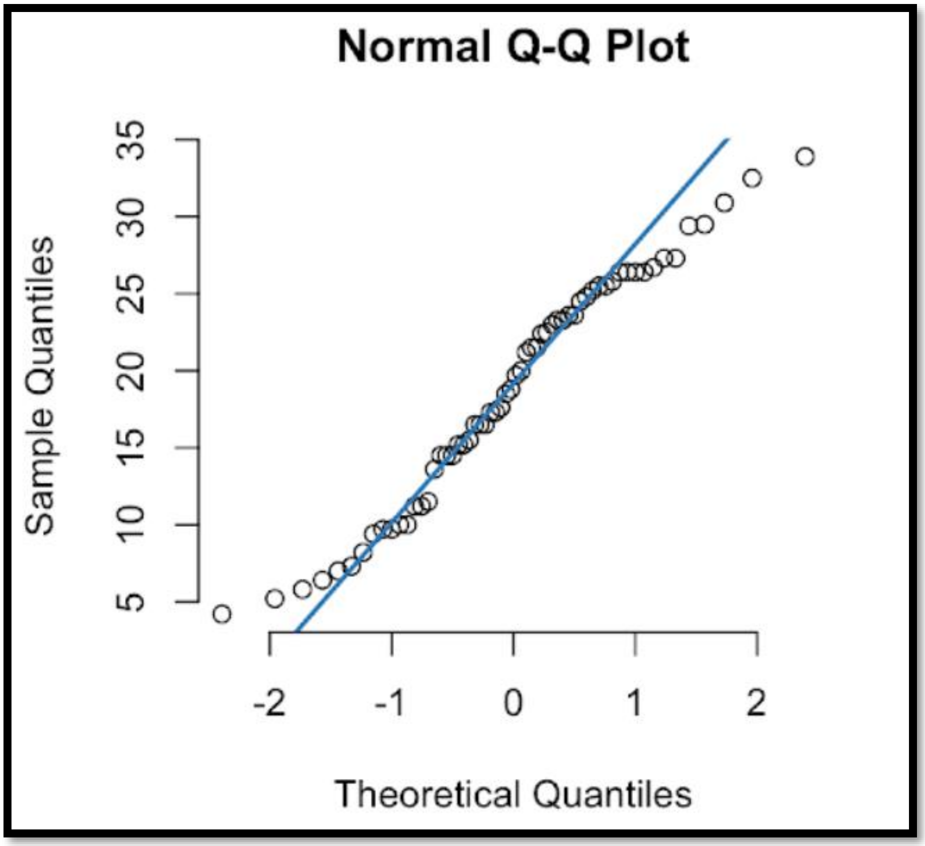
f′(m,b)=⎡⎣dfdmdfdb⎤⎦=[1N∑−2xi(yi−(mxi+b))1N∑−2(yi−(mxi+b))]f′(m,b)=[dfdmdfdb]=[1N∑−2xi(yi−(mxi+b)) 1N∑−2(yi−(mxi+b))]

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.

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

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.



We can also randomly generate data from a standard Normal distribution and then find the quantiles. Here we generate a sample of size 200 and find the quantiles for 0.01 to 0.99 using the quantile function:

```
quantile (rnorm (200), probs = seq(0.01,0.99,0.01))
```