1. Explain the linear regression algorithm in detail.
2. What are the assumptions of linear regression regarding residuals?
3. What is the coefficient of correlation and the coefficient of determination?
4. Explain the Anscombe’s quartet in detail.
5. What is Pearson’s R?
6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
8. What is the Gauss-Markov theorem?
9. Explain the gradient descent algorithm in detail.
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

SOLUTIONS:

1. Linear Regression is a machine learning algorithm based on the supervised learning. It basically predicts the linear relationship between the dependent variable/output variable(y) and the independent variable/input variable(x). We can calculate the (y) from the linear combination of the input variables (x).

When we want to find the linear relationship by using one input variable(x), it is called simple linear regression. When there are multiple input variables(x), then it is called multiple linear regression. To prepare or train the linear regression equation from the data, we used different technique but the most common is Ordinary Least Square(OLS).

The equation for simple linear regression:

y = B0 + B1\*X

Where, B0 = Slope and B1 = Intercept.

The equation for Multiple linear regression:

Y = B0 + B1\*X1 + B2\*X2…..Bp\*Xp

The OLS method has criterion of the minimization of the sum of the squares of the residuals. Residuals are defined as the difference between the y-coordinates of the actual data and y-coordinates of predicted data.

The main criterion used to determine the best-fitting regression line is the line that minimizes the sum of squares of distances of points from the regression line.

Best fit line is obtained by minimizing a quantity called Residual sum of squares(RSS). By achieving the best-fit regression line, the model aims to predict y value such that the difference between predicted value and true value is minimum. Cost function of linear regression is the Root Mean Squared Error ( RMSE ) between predicted y value and true y value. The following are the ways to minimize the cost function:

1. Differentiation.
2. Gradient Descent Approach.

We have a R square(R2):

R2 = 1 – RSS/TSS

RSS = Residual Sum of Squares

TSS = Total Sum of Squares.

The value of R2 always lies between 0 and 1. RSS with value 1 means the variance in the data is explained by the data, 0 means none of the variance is being explained by the model.

To scale all the numeric variables, we have MinMax and standardization scaling.

Standardization: x – mean(x)/SD(x)

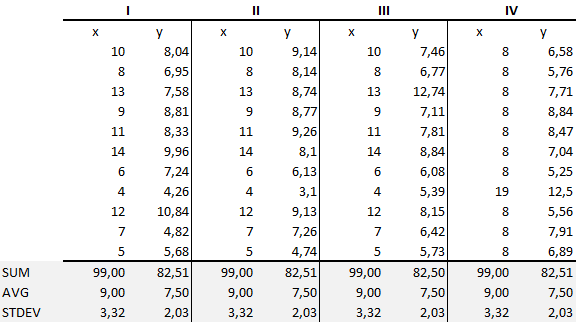
MinMax Scaling: x – min(x)/max(x) – min(x)

1. Assumptions of Linear Regression regarding residuals:
2. Residuals should be normally distributed.
3. Mean of the residuals should be zero.
4. Residuals should not be correlated with each other.
5. The variance of the residuals should be consistent for all the observations. It is known as homoscedasticity.
6. Multicollinearity should not be present in the residuals.

1. The coefficient of correlation measures the strength and linear relationship between two variables x and y. The value of correlation coefficient always lies between –1 and 1. If correlation is positive, it mean when x increases, y also increases. If correlation is negative, it means when x increases, y decreases. If correlation is zero or close to zero, it means there is no linear relationship between two variables. A perfect correlation occurs when the data points lie on a straight line.

The coefficient of determination(R2) measures how much variance of the data is explained by the model. It is basically the ratio of explained variation to the total variation. Its value always lies between 0 and 1. It measures how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variations. If the line is away from the points, it less explains it.

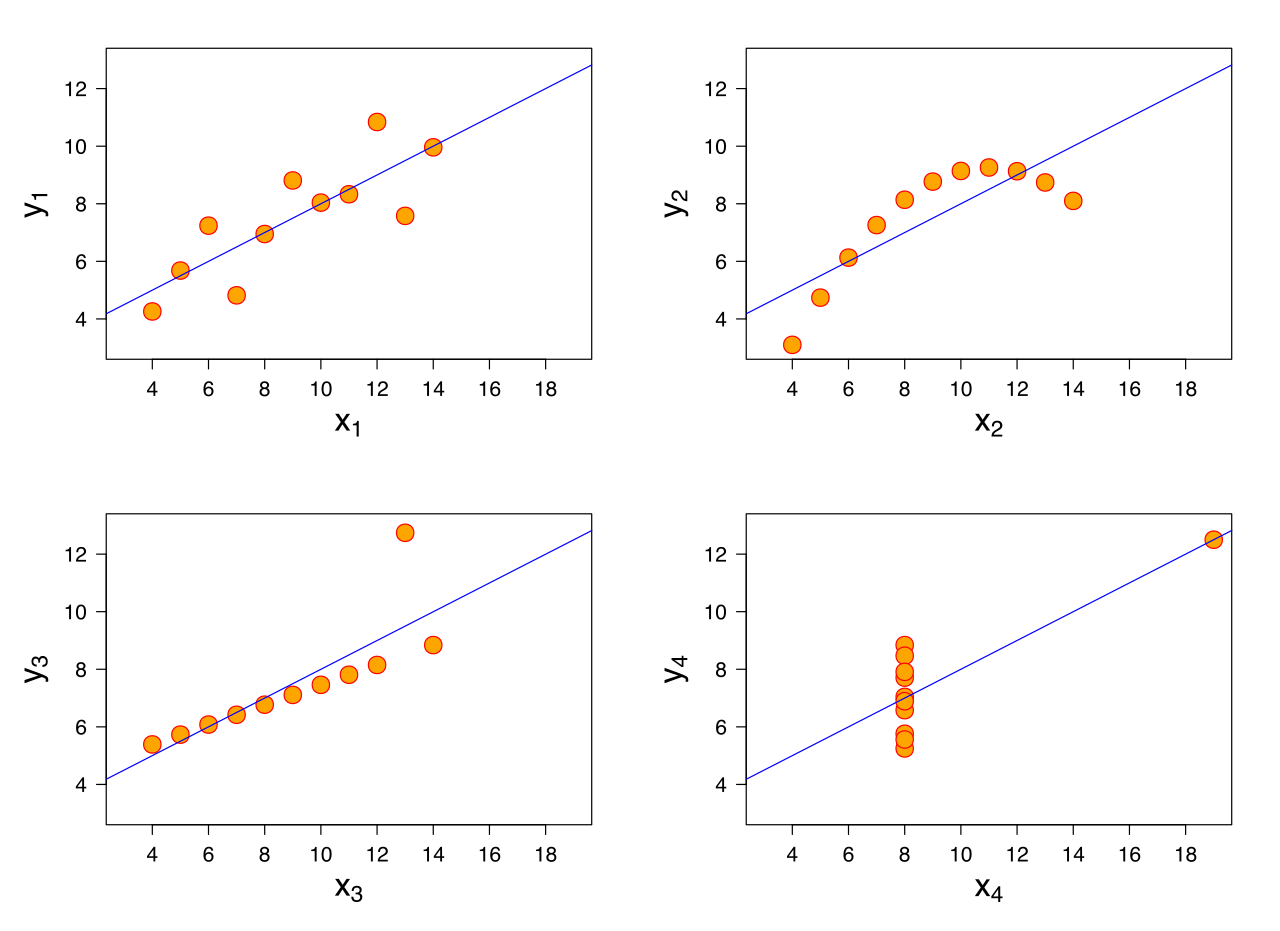
1. Anscombe’s Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed. Anscombe’s Quartet warns of the dangers of outliers in the data sets. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. Look at the following summary statistics:



From the summary statistics it is clear that mean and variance were identical for x and y across the groups:

1. Mean of x is 9 and mean of y is 7.50 in each of the four datasets.
2. Also, the variance of x is 11 and variance of y is 4.13 for each dataset.

Now, plot these 4 datasets and observe the patterns.



1. From Dataset 1, it is clear that it have clean and well-fitting linear models.
2. From Dataset 2, it shows non-normal distribution.
3. In Dataset 3, the distribution is linear.
4. In Dataset 4, high correlation coefficient is made due to the presence of only one outlier.

1. Pearson’s R is a statistic that measures linear correlation between two variables X and Y. It basically measures the strength of the association between the two variables .It has value between -1 and 1. The value of 1 implies that a linear equation describes the relationship between x and y perfectly, with all data points lying on a line with a positive slope for which y increases as x increases. 0 value means there is no linear correlation between two variables x and y. -1 means there is a negative slope and if x increases then y decreases .It is basically the covariance of the two variables divided by the product of their standard deviations. The t-test is used to establish if the correlation coefficient is significantly different from zero and hence that there is a evidence of an association between the two variables. If p-value is less then 5% significance level, we can say that the coefficient is significant ( p < 0.05 ). Requirements for Pearson’s correlation coefficient are:
2. There should not be outliers in the data.
3. The variables should be normally distributed.
4. The association should be linear.
5. Scaling is a technique to standardize the independent features present in the data. Basically we performed scaling during data pre processing. It basically scales all the values into the one unit. If we not perform scaling, then we consider higher values as large values and smaller values as the lower values, regardless of the unit of values. Also, we use feature scaling for the ease of interpretation and faster convergence for the gradient descent methods.
6. Standardized scaling: It brings all of the data into a standard normal distribution with mean zero and standard deviation one.

Standardization: x = x – mean(x) / SD(x)

1. MinMax Scaling: It brings all the data in the range of 0 and 1.

MinMax: x = x - min(x) / max(x) - min(x)

1. As we know that we use VIF to detect the multicollinearity. A high VIF means that specific variable is explained by some other variables. An infinite VIF values indicates that the corresponding variable may be expressed exactly by a linear combination of other variables, so that’s why some times it shows INF VIF value.
2. The Gauss Markov Theorem states that if a certain set of assumptions are met, the ordinary least squares (OLS) estimates for regression coefficients gives you the Best Linear Unbiased Estimate possible. Basically, the point of the Gauss Markov theorem is that we can find conditions ensuring a good fit without requiring detailed distributional assumptions about the error terms and without distributional assumptions about the independent variables. There are following some Gauss Markov Assumptions:
   1. Non-Collinearity: The regressors which we calculated aren’t perfectly correlated with each other.
   2. Homoscedasticity: The variance of the Error terms must be constant.
   3. Linearity: Parameters must be linear which we estimates using OLS ( Ordinary Least Square).

If all the Gauss-Markov assumptions are met than the OLS estimators alpha and beta are BLUE( Best Linear Unbiased Estimators )

Best means Variance of the OLS estimators is minimal, smaller than the variance of any other estimators.

Linear means if relationship is not linear OLS is not applicable.

Unbiased means the expected values of the estimated beta and alpha equal the true values which describes the relationship between x and y.

1. Gradient Descent is an algorithm that is used to minimize a function. These algorithms achieve this end by starting with initial parameter values and iteratively moving towards a set of parameter values that minimize some cost function. The movement toward best fit is achieved by taking the derivative of the single or more variables which involved in it, towards the direction which is having lowest gradient. Ordinary Linear regression is a simple way of explaining how gradient descent works. Our main objective in the gradient descent is that we have to minimize the values of actual y and values of predicted y. Following is the code to calculate the gradient descent. We implement this by using MSE (Mean Squared Error ) . MSE is RSS divided by number of points (n).

def \_\_init\_\_(self, rate=0.01, iter=1000):

self.rate, self.iter = rate, iter

def fit(self, X, y):

c = 0

m = 5

n = X.shape[0]

for i in range(self.iter):

c\_gradient = -2 \* np.sum(y - m\*X + c) / n

m\_gradient = -2 \* np.sum(X\*(y - (m\*X + c))) / n

c = c + (self.rate \* c\_gradient)

m = m - (self.rate \* m\_gradient)

self.m, self.c = m, c

def predict(self, X):

return self.m\*X + self.c

Where m is the slope of the line of best fit, c is the intercept.

1. Q-Q Plots ( Quantile-Quantile Plots ) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. Lets take an example of median in which 50% of the data fall below that point and 50% lie above it. The main purpose of the Q-Q plot is to find out if two sets of data comes from the same distribution. A 45 degree angle is plotted on the Q-Q plot, if two data sets come from a common distribution, the points will fall on that reference line. This particular type of Q-Q plot is called a normal Quantile-Quantile Plot. The Q-Q Plot compares the quantiles of our data against the quantiles of the desired distribution.

Steps to make a Q-Q plot:

1. Order the items from smallest to largest.
2. Draw a normal distribution curve.
3. Find the z-value fo each segment.
4. Plot your data set values against your normal distribution z-value.

Advantages of using Q-Q Plot:

1. We can used it with the sample sizes also.
2. Many distributional aspects like shifts in scale, shifts in symmetry and the presence of outliers can all be detected from this plot.

Some possible interpretations if we have two datasets:

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x-axis.
2. Y < X: It means if y-quantiles are lower than the x-quantiles.
3. X < Y: It means if x-quantiles are lower than the y-quantiles.
4. Different Distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x-axis.