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.

Answers:

(searched internet for the answers)

Explain the linear regression algorithm in detail.

Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as:

$Y \approx \beta 0 + \beta 1X.$

For example, X may represent TV advertising and Y may represent sales. Then we can regress sales onto TV by fitting the model:

$sales \approx \beta 0 + \beta 1 \times TV.$

β0 and β1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, β0 and β1 are known as the model coefficients or parameters. Once we have used our training data to produce estimates $\hat{}\ \beta 0$ and $\hat{}\beta 1$ for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising by computing

$\hat{}y = \hat{}\ \beta 0 + \hat{}\ \beta 1x$

where $\hat{}y$ indicates a prediction of Y on the basis of X = x. Here we use a hat symbol, $\hat{}$, to denote the estimated value for an unknown parameter or coefficient, or to denote the predicted value of the response.

What are the assumptions of linear regression regarding residuals?

The assumptions of simple linear regression are:

1. Linear relationship between X and Y

2. Error terms are normally distributed (not X, Y)

3. Error terms are independent of each other

4. Error terms have constant variance (homoscedasticity)

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

The quantity r, called the linear *correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The mathematical formula for computing ***r*** is:

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

The value of r is such that -1 < r < +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

Positive correlation: 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.

Negative correlation: 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.

No correlation: 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 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.

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. The coefficient of determination is such that 0 < r^2 < 1, and denotes the strength of the linear association between x and y. The coefficient of determination represents the percent of the data that is the closest to the line of best fit.

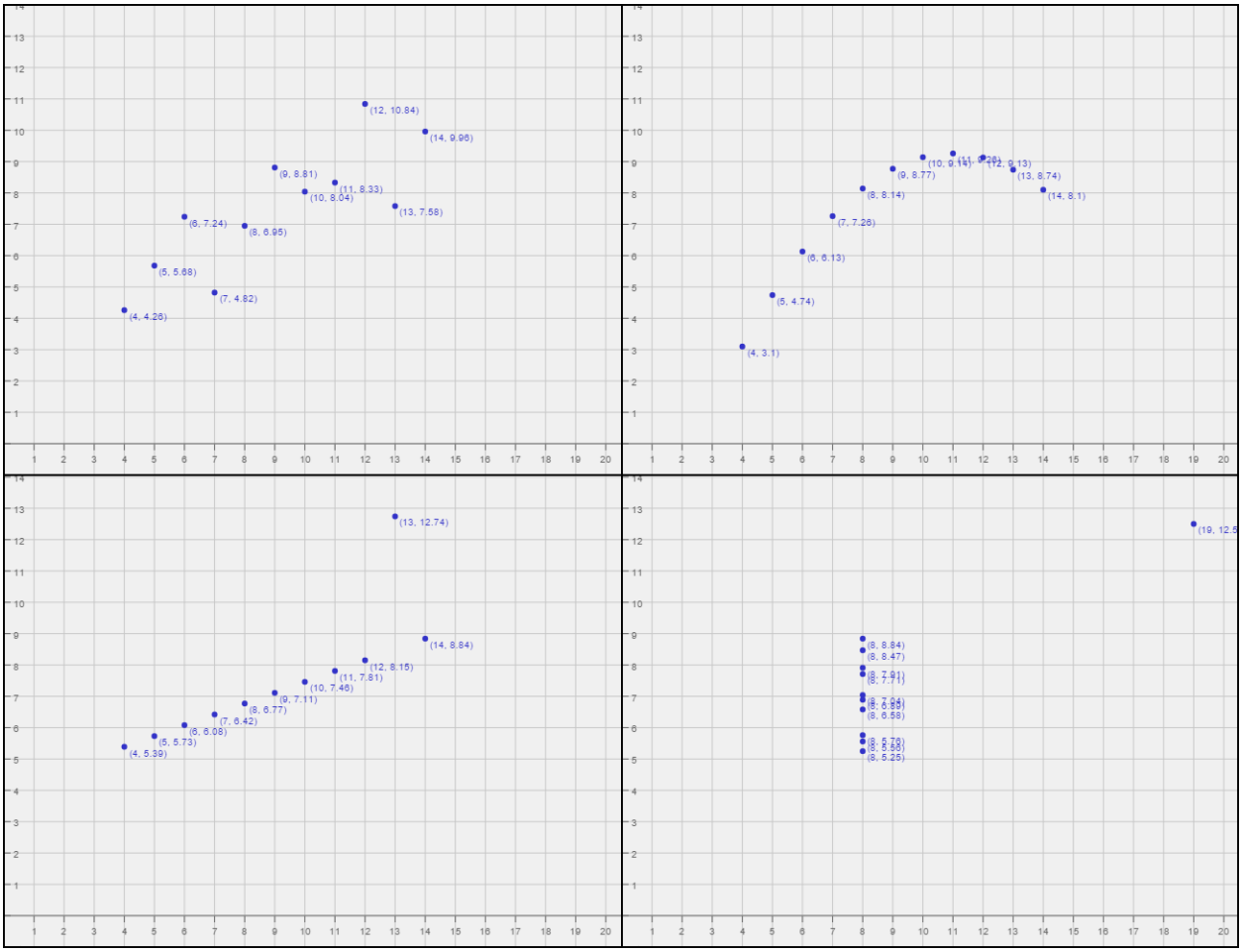
For example, if r = 0.922, then r^2 = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

RSS = Residual sum of square, TSS = Sum of errors of the data from mean

Explain the Anscombe's quartet in detail.

Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as "Anscombe's Quartet," are shown below.



All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. it comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

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 is still 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

What is Pearson's R?

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

For a population:

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables (X,Y) the formula for ρ is:

$$P_{X,Y} = \text{cov}(X,Y) / \rho_X \ \rho_Y$$

Where cov is the covariance, ρ_X is the standard deviation of X, ρ_Y is the standard deviation of Y.

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

When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

Having features on a similar scale can help the gradient descent converge more quickly towards the minima – that is why scaling is performed.

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.

$$X_{changed}=(X-X_{min}) / (X_{max}-X_{min})$$

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

$$X_{changed} = (X-\mu) / \sigma$$

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. If two Xs are perfectly correlated then

VIF = 1/(1-1)= 1/0 = infinity

What is the Gauss-Markov theorem?

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives 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.

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.

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

$cov\{\varepsilon i, \varepsilon j\} = 0, i, j = 1,\ldots., N\ I \neq j.$

$V\{\varepsilon 1 = \sigma 2, i= 1, \ldots.N$

Explain the gradient descent algorithm in detail.

Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. To find a local minimum of a function using gradient descent, we take steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point. But if we instead take steps proportional to the positive of the gradient, we approach a local maximum of that function; the procedure is then known as gradient ascent. Gradient descent is also known as steepest descent.

Gradient descent is an optimization algorithm that finds the optimal weights (a,b) that reduces prediction error, the Gradient Descent algorithm stated step by step as follows:

Step 1: Initialize the weights(a & b) with random values and calculate Error (SSE)

Step 2: Calculate the gradient i.e. change in SSE when the weights (a & b) are changed by a very small value from their original randomly initialized value. This helps us move the values of a & b in the direction in which SSE is minimized.

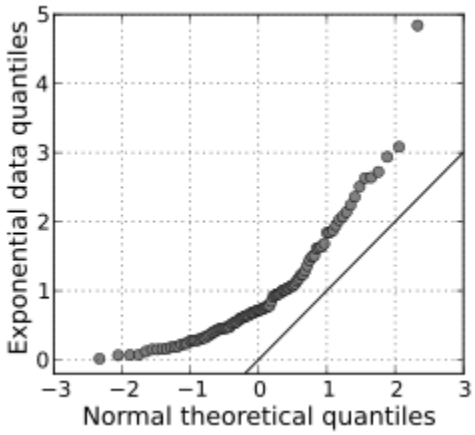
Step 3: Adjust the weights with the gradients to reach the optimal values where SSE is minimized

Step 4: Use the new weights for prediction and to calculate the new SSE

Step 5: Repeat steps 2 and 3 till further adjustments to weights doesn't significantly reduce the Error

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

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. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



A Q Q plot showing the 45 degree reference