Question [1]. Explain the linear regression algorithm in detail.

Answer:

Linear regression algorithm

Linear Regression is an algorithm based supervised machine learning where the predicted variable is continuous and has a constant steep slope. It's used to predict variables within a continuous range, (e.g. sales) rather than trying to classify them into categories.

Linear regression is a **linear model**, i.e. a model that establishes a linear relationship between the independent variables (x) and the one dependent variable (y) (using a **best fit straight line** also known as regression line). More specifically, that y can be calculated from a linear combination of the independent variables (x). Independent variables are also called 'Predictors'.

It is represented by an equation Y= β0 + β1*X + e, where β0 is intercept, β1 is slope of the line and e is error term. This equation can be used to predict the value of target variable based on predictor variable(s).

Regression analysis is widely used for 2 purposes: **a) Forecasting and b) Prediction**. The uses of forecasting and prediction have substantial overlap. However, they are different and it's important to understand why, to be able to use regression effectively for each purpose.

There are two main types, when there is a one independent variable (x), the method is referred to as **simple linear regression**. When there are multiple independent variables, literature from statistics often refers to the method as **multiple linear regression.**

Form of Linear Regression

Mathematically, we can write a linear relationship as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Where:

- ***y*** is the response
- ***β*** values are called the **model coefficients**.
- ***β***0 is the intercept
- ***β***1 is the coefficient for *X1* (the first variable)
- ***βn*** is the coefficient for *Xn* (the nth variable)

Linear Regression – Learning the Model

Simple Linear Regression

With simple linear regression when we have a single input, we can use statistics to estimate the coefficients. This requires to calculate statistical properties from the data such as mean, standard deviation, correlation, and covariance.

Ordinary Least Squares

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients. Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals.

Gradient Descent

This operation is called Gradient Descent and works by starting with random values for each coefficient. The sum of the squared errors is calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

Regularization

There are extensions to the training of the linear model called regularization methods. These seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model (like the number or absolute size of the sum of all coefficients in the model).

Question [2]. What are the assumptions of linear regression regarding residuals?

Answer:

Assumptions of linear regression regarding residuals

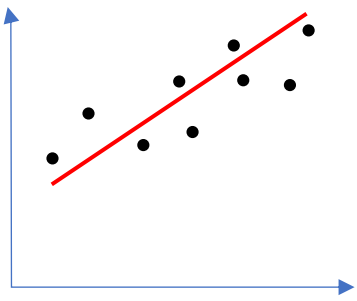
- The mean of residuals is zero.
- Homoscedasticity of residuals or equal variance.
- Normality of residuals.
- No autocorrelation of residuals: This is applicable especially for time series data. Autocorrelation is the correlation of a time Series with lags of itself. When the residuals are autocorrelated, it means that the current value is dependent of the previous (historic) values and that there is a definite unexplained pattern in the Y variable that shows up in the disturbances.
- The X variables and residuals are uncorrelated.

Question [3]. What is the coefficient of correlation and the coefficient of determination?

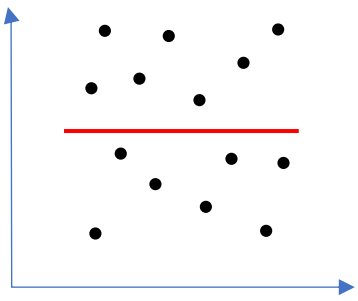
Answer:

Coefficient of correlation

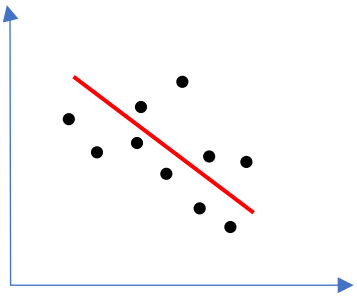
The correlation coefficient is a statistical measure that calculates the strength of the relationship between the of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1. 0 indicates a strong negative relationship, while a correlation of 1.0 indicates a strong positive relationship. A correlation of 0.0 shows no relationship at all between the variables.



Positive Correlation



No Correlation



Negative Correlation

Coefficient of determination

The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1. The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor.

The closer the value is to 1, the better the fit, or relationship, between the two factors. The coefficient of determination is the square of the correlation coefficient, also known as "R".

- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

For a model with several variables, such as a multiple regression model, the adjusted R^2 is a better coefficient of determination. In economics, an R^2 value above 0.60 is seen as worthwhile.

Mathematically, it is represented as: $\mathbf{R^2 = 1 - (RSS / TSS)}$, where **RSS** (Residual Sum of Squares) and **TSS** (Total sum of squares)

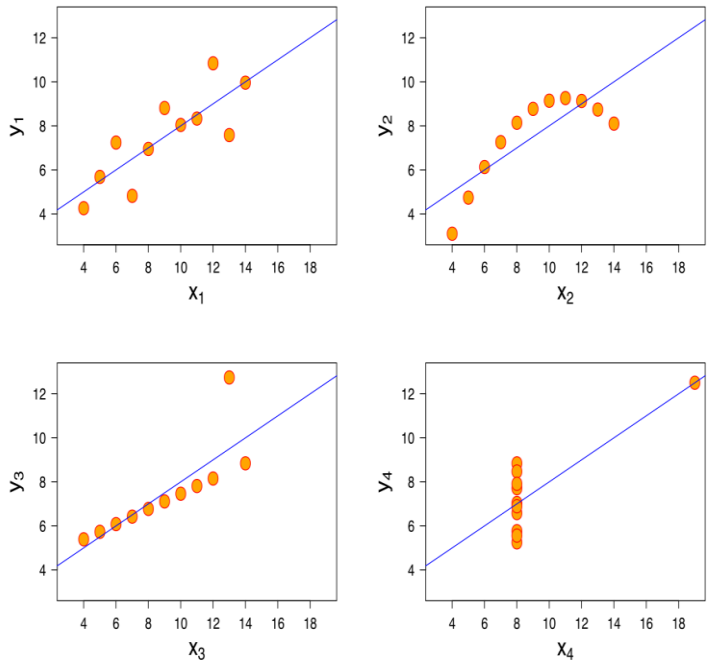
The coefficient of determination for (R^2) a linear regression model with one independent variable is:

$$\mathbf{R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y) \}^2}$$

Question [4]. Explain the Anscombe's quartet in detail.

Answer:

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) 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 first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant. In the third graph (bottom left), the distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear

All four datasets:

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Properties of all four datasets:

- **Mean** of ***x*** in each case: **9** (exact)
- **Variance** of ***x*** in each case: **11** (exact)
- **Mean** of ***y*** in each case: **7.50** (to 2 decimal places)
- **Variance** of ***y*** in each case**: 4.122** or **4.127** (to 3 decimal places)
- **Correlation** between ***x*** and ***y*** in each case: **0.816** (to 3 decimal places)
- **Linear regression line** in each case: ***y* = 3.00 + 0.500*x*** (to 2 and 3 decimal places, respectively)

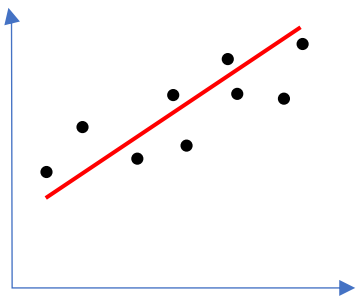
Question [5]. What is Pearson's R?

Answer:

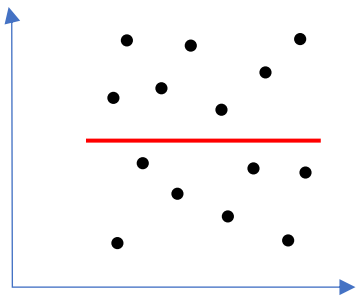
Pearson's R

Pearson's *r* which is also known as **Pearson correlation coefficient** (**PCC)** is a measure of the linear correlation between two variables *X* and *Y*. Its value ranges between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. 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.

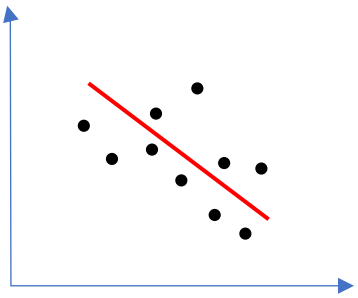
The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample.



Positive Correlation



No Correlation



Negative Correlation

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

Answer:

Scaling

It is a step of Data Pre-processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Real world dataset contains features that highly vary in magnitudes, units, and range. Scaling should be performed when the scale of a feature is irrelevant or misleading. When you 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. We need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

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.

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

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

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

For most application standardization is recommended.

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

Answer:

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If the VIF for a variable is 16 the associated standard error is four times as large as it would be if its VIF was 1. In such a case, the coefficient would have to be 4 times as large to be statistically significant at a given significance the level.
The VIF can be conceived as related to the R-squared of a particular predictor variable regressed on all other includes predictor variables:

$$VIF = \frac{1}{1 - R^2}$$

If two Variables are perfectly correlated or When R-squared reaches 1, VIF reaches infinity

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

that is the estimate is as imprecise as it can be.

Question [8]. What is the Gauss-Markov theorem?

Answer:

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 you the best linear unbiased estimate (BLUE) possible.

Gauss Markov Assumptions

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

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.

Question [9]. Explain the gradient descent algorithm in detail.

Answer:

Gradient descent algorithm

Gradient descent is an optimization algorithm that optimize the objective function (cost function for linear regression) to reach the optimal solution.

Or

Gradient descent is an optimization algorithm used to find the values of the parameters (coefficients) of a function (f) that minimizes a given cost function (cost).

It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function w.r.t the parameters where the gradient gives the direction of the steepest ascent. The size of the step we take on each iteration to reach the local minimum is determined by the learning rate α. Therefore, we follow the direction of the slope downhill until we reach a local minimum.

The equation for the line that's fit the data, is given as:

$$y(p) = \theta_0 + \theta_1 X$$

Where θ_0 is the intercept of the fitted line and θ_1 is the coefficient for the independent variable x

To find the optimum thetas, we need to reduce the cost function for all data points, which is given as,

$$J(\theta_0, \theta_1) = \sum_{i=1}^{N} (y_i - y_i(p))^2$$

The way to find the optimal thetas is known as Gradient Descent. There are two types of Optimization methods,

- Closed form solution
- Iterative form solution

Gradient descent is an iterative form solution of order one. So as to compute optimal thetas, we need to apply Gradient Descent to the Cost function, which is given as follows,

$$\frac{\partial}{\partial\theta}J(\theta)$$

To compute θ_1, the equation will look like this,

$$\theta^1 = \theta^0 - \eta\frac{\partial}{\partial\theta}J(\theta)$$

Where η is known as the learning rate, which defines the speed at which we want to move towards negative of the gradient. (learning rate controls how much the coefficients can change on each update.)

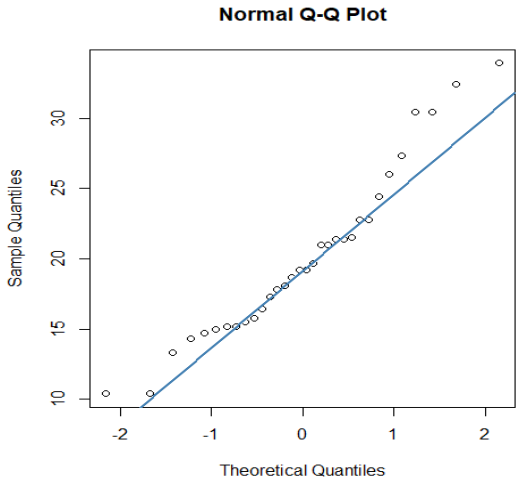
This process is repeated until the cost of the coefficients is 0.0 or close enough to zero to be good enough.

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

Answer:

Q-Q plot

The quantile-quantile or **q-q plot** is an exploratory graphical device **used to** check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question.



The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Use a Q-Q plot in linear regression

- The Q-Q plot can also be used to compare two distributions based on a sample from each. If the samples are the same size then this is just a plot of the ordered sample values against each other. Choosing a fixed set of quantiles allows samples of unequal size to be compared.
- The Q-Q plot is used to determine if two data sets come from populations with a common distribution.
- The Q-Q plot is used to determine if two data sets have common location and scale.
- The Q-Q plot is used to determine if two data sets have similar tail behaviour.
- Q-Q plots can be used to assess how well a theoretical family of models fits your data, or your residuals.
- The Q-Q plot can used to check the validity of a distributional assumption for a data set.
- Q-Q plots are helpful for understanding departures from a theoretical model.
- When normality is not proved, the presence of outliers is often the cause; the Q-Q plot is then recommended for detection of influential points.

Importance of a Q-Q plot

- When the Histogram is not useful for evaluating the fit of the chosen distribution. When there are a small number of data points, a histogram can be rather ragged. Further, our understanding of the fit depends on the widths of the histogram intervals. But even if the intervals are well chosen, grouping data into cells makes it difficult to compare a histogram to a continuous probability density function.
- The Q-Q Plots or Quantile-Quantile Plots overcomes all the limitations of the Histogram plot.
- Both empirical procedures, the Histogram & Q-Q plots, can identify the general form of the distribution. But among these two, Q-Q plots perform well even for small sample size. As most of our statistical methods depend on the normality assumptions, checking normality for the sample data is important. If we violate this assumption then the inference driven from the analysis may not be precise.
- Check for Common Distribution: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.