1. In regression analysis, which of the statements is true?

1. The mean of residuals is always equal to Zero
2. The Mean of residuals is less than Zero at all times
3. The Mean of residuals is more than Zero at all times
4. You do not have any such rule for residuals.

The correct answer is A. In regression analysis, the sum of the residuals in regression is always equal to Zero. Thus, it implies that the mean will also be Zero if the sum of the residuals is Zero.

Which of the statements is correct about Heteroscedasticity?

1. Linear regression with different error terms
2. Linear regression with constant error terms
3. Linear regression with no error terms
4. None of the above

The solution is the option A. When you have a non-constant variance in the error terms, it results in Heteroscedasticity. Such non-constant variance occurs because you have outliers.

Which of the following plots is best suited to test the linear relationship of independent and dependent continuous variables?

1. Scatter Plot
2. Bar Chart
3. Histograms
4. None of the above options

The answer is A. The Scatter plot is the best way to determine the relationship between continuous variables. You can find out how one variable changes with respect to the other.

If you have only one independent variable, how many coefficients will you require to estimate in a simple linear regression model?

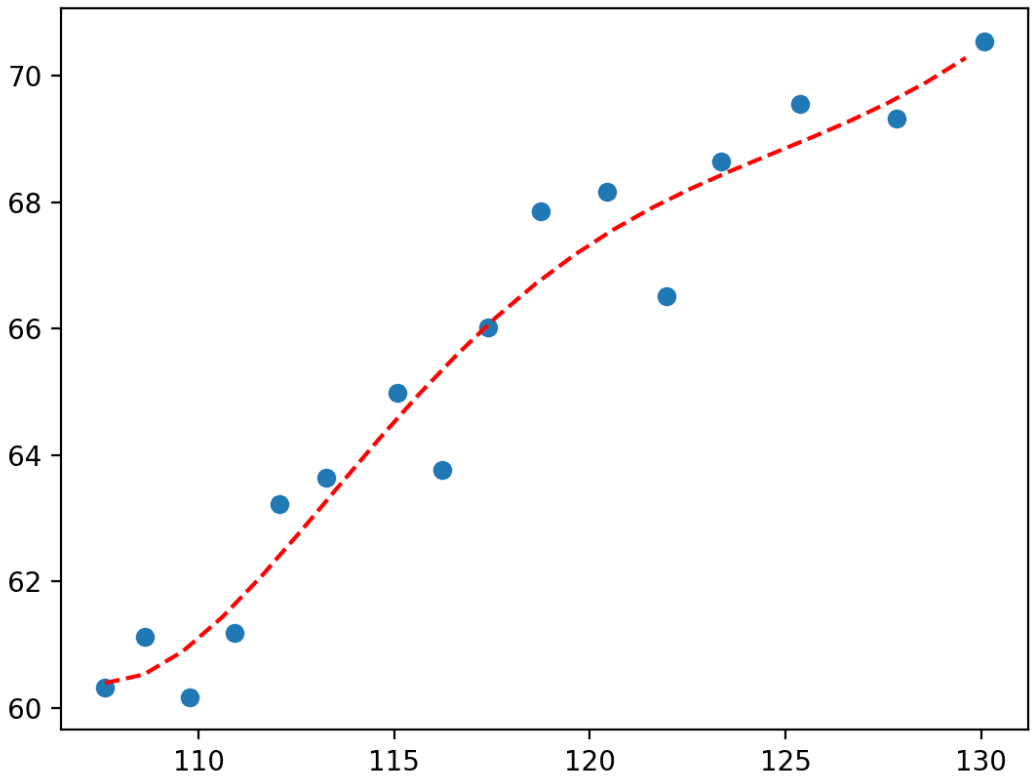
1. One
2. Two
3. No idea

The answer is B. Consider the simple linear regression with one independent variable. Y = a + bx. You can see that you need two coefficients.

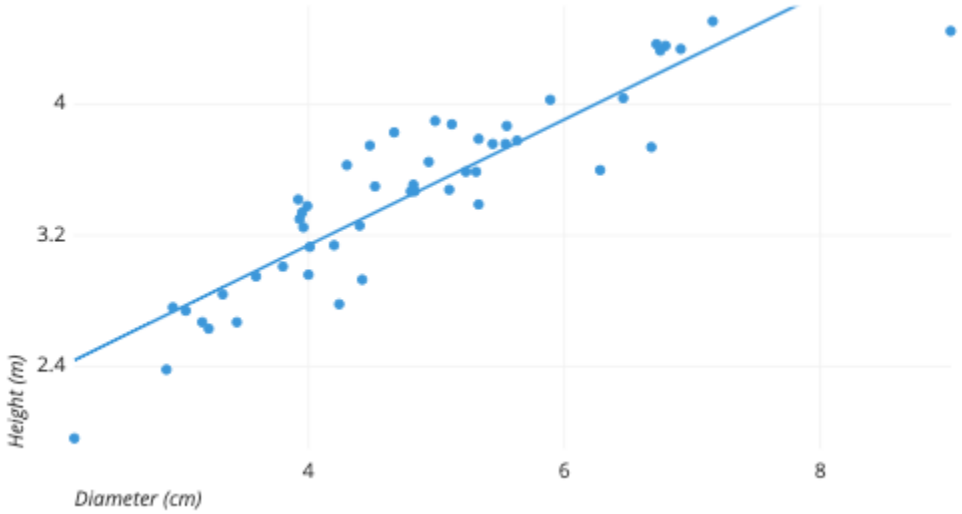
How does a *Non-Linear* regression analysis differ from *Linear* regression analysis?

- ***Non-linear*** functions have variables with powers greater than 1. Like x^2*x*2. If these non-linear functions are graphed, they do not produce a straight line (their direction changes constantly).
- ***Linear*** functions have variables with only powers of 1. They form a straight line if it is graphed.

- ***Non-linear*** regression analysis tries to model a non-linear relationship between the independent and dependent variables.
- A simple non-linear relationship is shown below:



- ***Linear*** regression analysis tries to model a linear relationship between the independent and dependent variables.
- A simple linear relationship is shown below:



How is the *Error* calculated in a Linear Regression model?

Junior

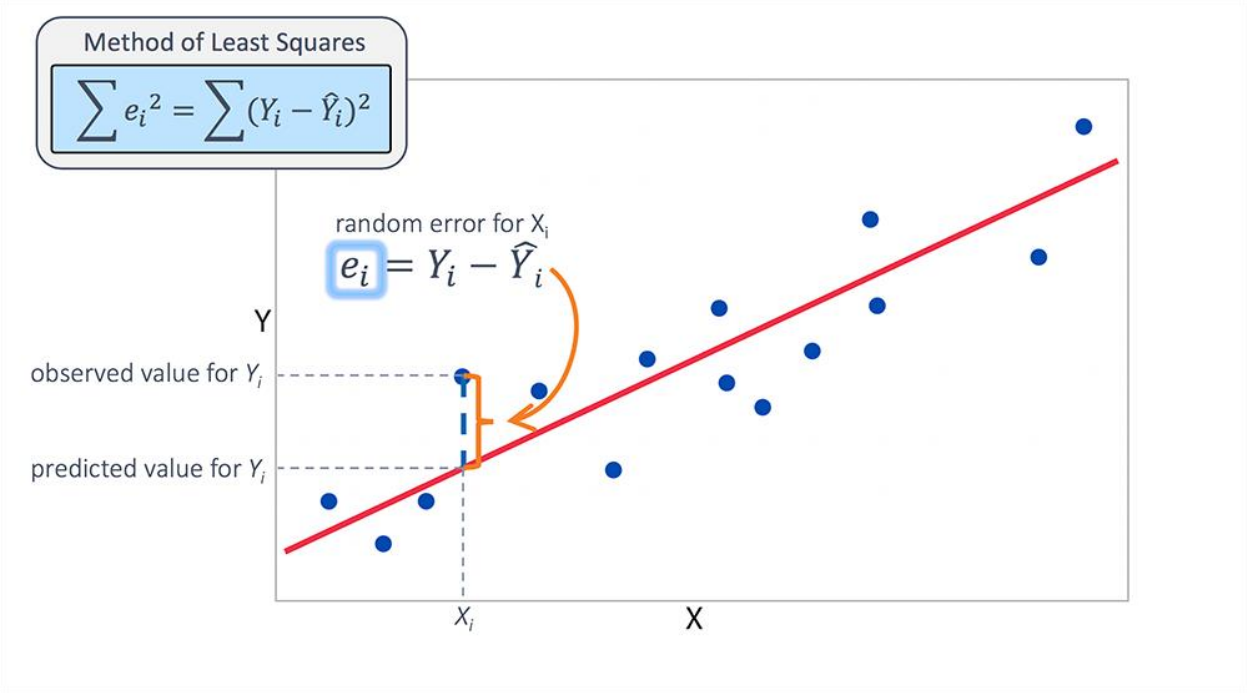
Linear Regression 48

Answer

1. Measuring the distance of the observed *y-values* from the predicted *y-values* at each value of *x*.
2. Squaring each of these distances.
3. Calculating the *mean* of each of the squared distances.

MSE = (1/n) * Σ(actual – forecast)2

1. The smaller the **Mean Squared Error**, the closer you are to finding the *line of best fit*
2. How *bad* or *good* is this final value always depends on the context of the problem, but the main goal is that its value is so minimal as possible.



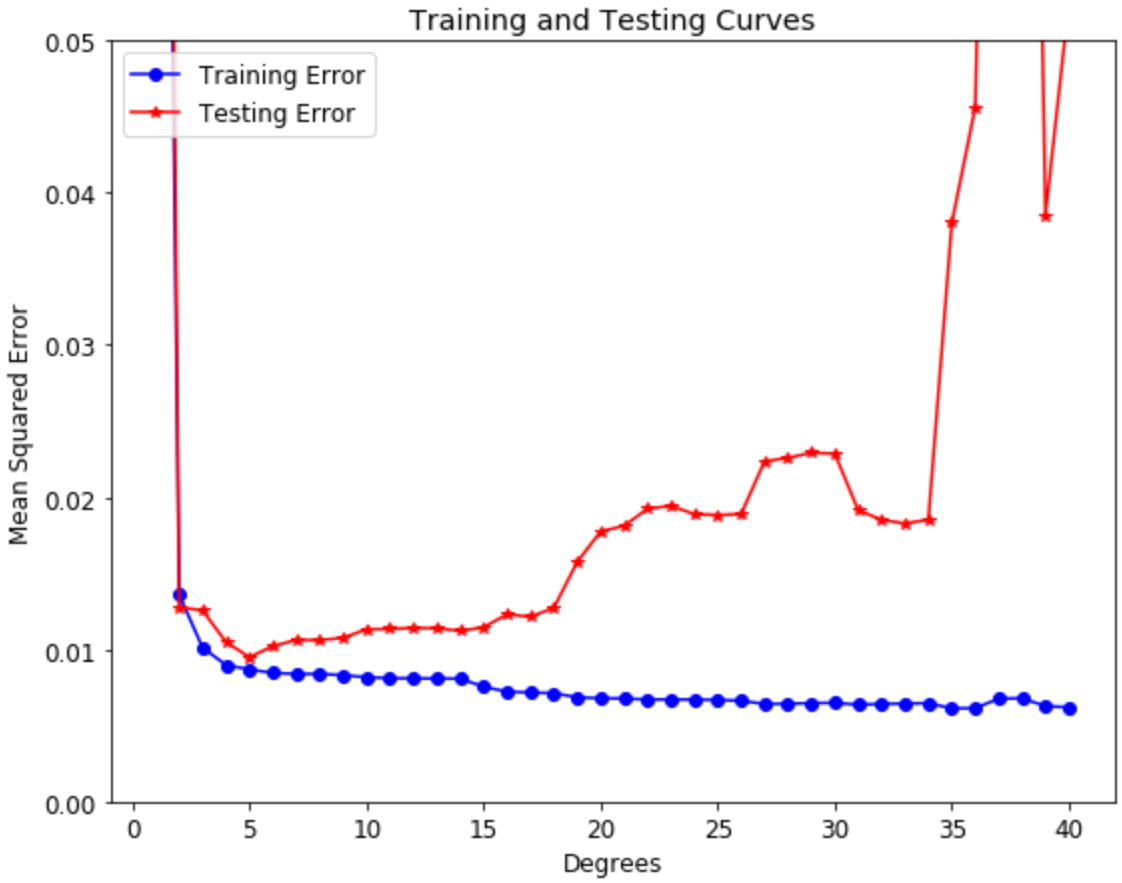
How would you detect *Overfitting* in Linear Models?

Junior

Linear Regression 48

Answer

The common pattern for **overfitting** can be seen on **learning curve** plots, where model performance on the training dataset continues to improve (e.g. loss or error continues to fall) and performance on the test or validation set improves to a point and then begins to get worse.



So an overfit model will have **extremely low training error but a high testing error**.

What are *types* of Linear Regression?

Junior

Linear Regression 48

Answer

- Simple **linear** regression uses traditional slope-intercept form. x represents our input data and y represents our prediction.

  $y = mx+b$

- A more complex, **multi-variable** linear equation might look like this, where *w* represents the coefficients, or weights, our model will try to learn.

  $f(x,y,z) = w1x+w2y+w3z$

  The variables x, y, z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

  $Sales = w1Radio+ w2TV+w3*News$

What is the difference between *Mean Absolute Error (MAE)* vs *Mean Squared Error (MSE)*?

Junior

Linear Regression 48

Answer

- The **Mean Squared Error** measures the variance of the residuals and is used when we want to punish the outliers in the dataset. It's defined as:
- $MSE=N1\sum(yi-y\hat{}\,)2$
- The **Mean Absolute Error** measures the average of the residuals in the dataset. Is used when we don't want outliers to play a big role. It can also be useful if we know that our distribution is multimodal, and it's desirable to have predictions at one of the modes, rather than at the mean of them. It's defined as:

$MAE=N1\sum|yi-y\hat{}\,|$

Compare *Linear Regression* and *Decision Trees*

Mid

Decision Trees 47

Answer

- **Linear regression** is used to predict *continuous* outputs where there is a linear relationship between the features of the dataset and the output variable.
- **Decision trees** work by splitting the dataset, in a tree-like structure, into smaller and smaller subsets and make predictions based on which subset the new example falls into.

- **Linear regression** is used for *regression* problems where it predicts something with infinite possible answers such as the price of a house.
- **Decision trees** can be used to predict both *regression* and *classification* problems.

- **Linear regression** is prone to *underfitting* the data. Switching to *polynomial regression* will sometimes help in countering underfitting.
- **Decision trees** are prone to *overfit* the data. *Pruning* helps with the overfitting problem.

Explain how does the *Gradient descent* work in *Linear Regression*

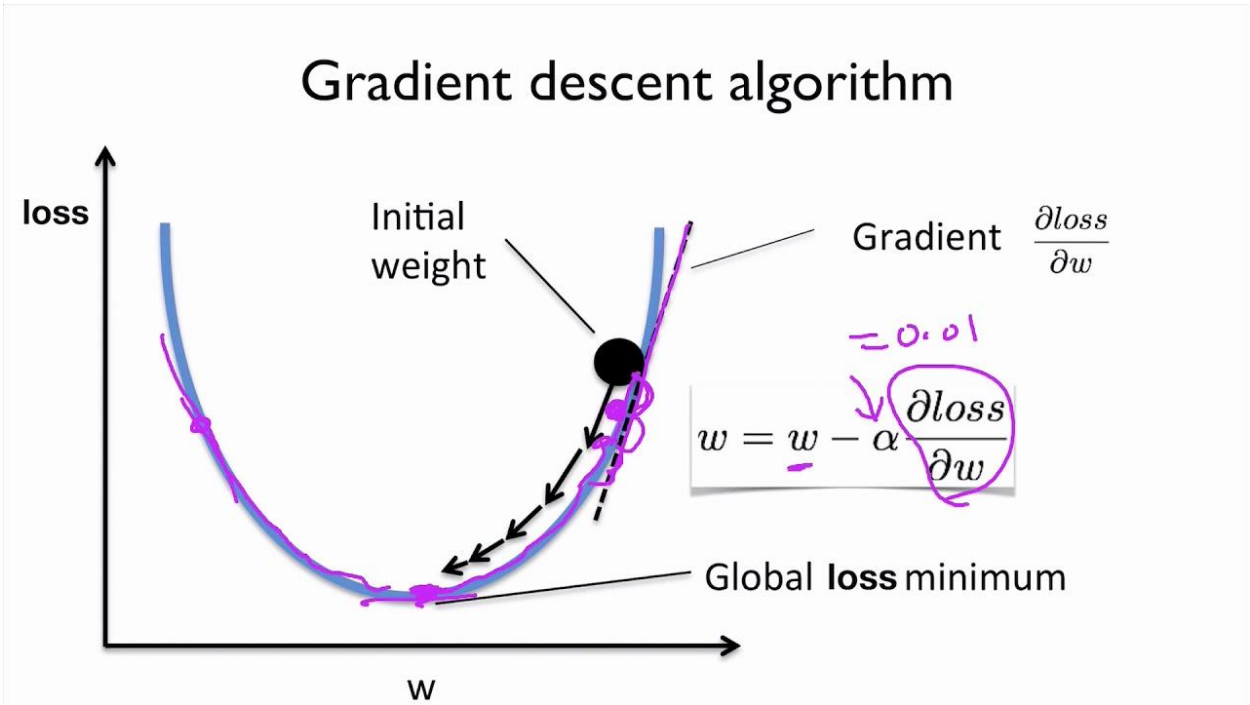
Mid

Linear Regression 48

Answer

The **Gradient Descent** works by starting with random values for each coefficient in the linear regression model.

- After this, the *sum of the squared errors* is calculated for each pair of input and output values (loss function), using a *learning rate* as a scale factor.
- For each iteration, the coefficients are updated in the direction towards *minimizing the error*,
- then we keep repeating the iteration process until a *minimum sum squared error* is achieved or no further improvement is possible.



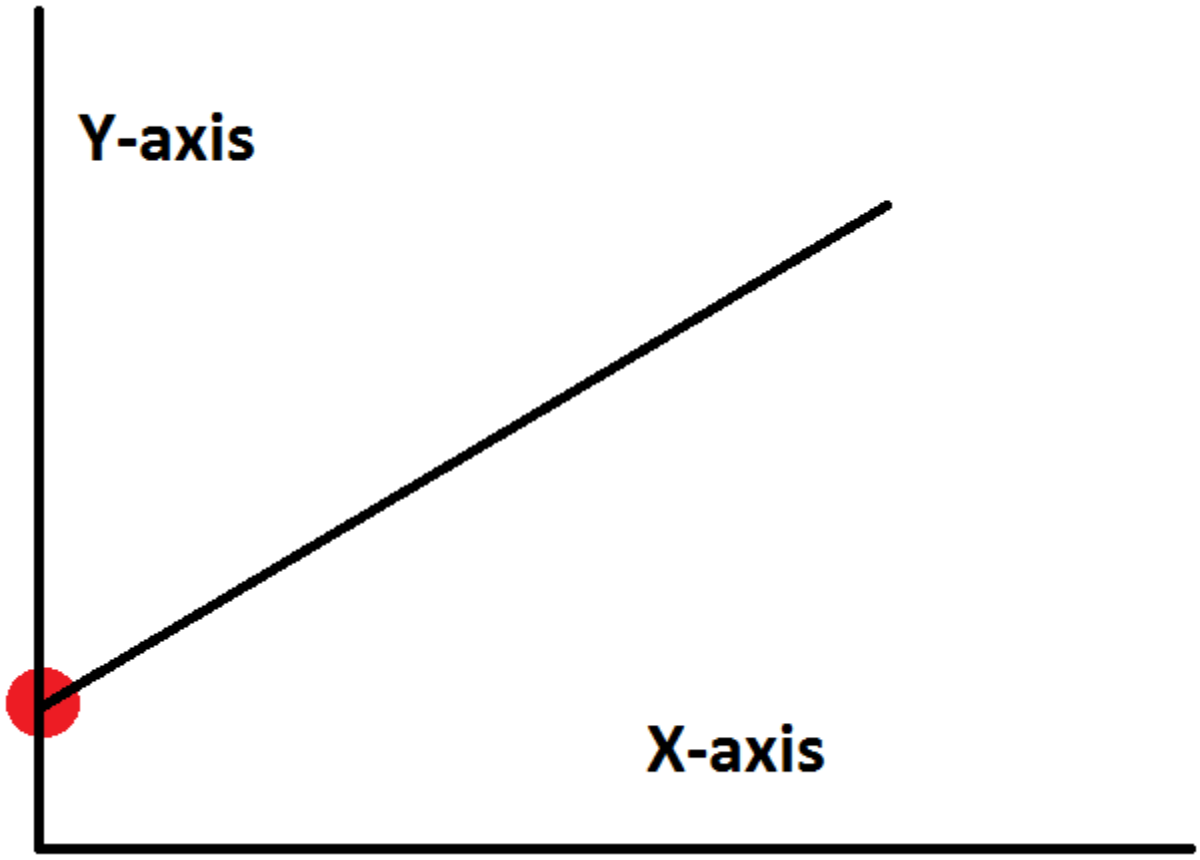
Explain what the *Intercept Term* means

Mid

Linear Regression 48

Answer

The constant term in regression analysis is the value at which the regression line crosses the y-axis. The constant is also known as the **y-intercept**.



The **intercept term** signifies the independent variable's shift from the origin and ensures that the model would be **unbiased**.

If we omit the intercept term, then the model is forced to go through the origin and the slope would become steeper and biased. Hence, we should not remove the intercept term unless we are completely sure that it is 0 according to theory and expectations.

How is *Hypothesis Testing* using in *Linear Regression*?

Mid

Linear Regression 48

Answer

Hypothesis testing is used in a *linear regression* model to test if the β1 parameter in the linear equation is *statistically significant*. In other words, to check if the linear relationship that we obtained was not caused just by random chance.
The way to use *hypothesis testing* is described as follows:

1. We start establishing the *null hypothesis (H0)* that β1 is not significant, i.e. there is no linear relationship between independent variables and the dependent variable. The *alternative hypothesis (H1)* is then that β1 is not zero.
    - H0 : β1=0
    - HA : β1≠0

1. We compute the *test statistic* which could it be the *T-test* or the *Z-test* depending on how many samples the dataset has.
2. We compute the corresponding *p-value*.
3. If the *p-value* turns out to be less than 0.05, we can reject the *null hypothesis* and state that β1 is indeed significant at the 95% confidence level.

With this, we can validate that our model coefficients are not obtained just by random chance.

How would you decide on the *importance* of variables for the *Multivariate Regression* model?

Mid

Linear Regression 48

Answer

A way to perform the **variable selection** is trying out different models, each containing a different subset of the predictors. For instance, if the number of predictors is 2, then we can consider 4 models:

1. A model containing no variables.
2. A model containing X1 only.
3. A model containing X2 only.
4. A model containing both X1 and X2.

We can then select the best model out of all of the models that we have considered by computing some statistics like **Adjusted R-squared**. However, if the number of predictors is high, we must use some more elaborated methods for feature selection, like:

- **stepwise regression**,
- **forward selection**, and
- **backward elimination**.

Name a disadvantage of *R-squared* and explain how would you address it?
Mid

Linear Regression 48

Answer

R-squared (R^2) is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

R-squared takes values between 0 and 1, with 0 indicating that the proposed model does not improve prediction over the mean model and 1 indicating the perfect prediction. However, one **drawback** of R-squared is that its values can increase if we add predictors to the regression model, leading to a possible *overfitting*.
To address this issue, we can use **Adjusted R-squared**: a modified version of *R-squared* that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance, and it decreases when a predictor improves the model by less than expected.

What is the difference between R square and adjusted R square?

R square and adjusted R square values are used for model validation in case of linear regression. R square indicates the variation of all the independent variables on the dependent variable. i.e. it considers all the independent variable to explain the variation. In the case of Adjusted R squared, it considers only significant

variables(P values less than 0.05) to indicate the percentage of variation in the model.

Name some *Evaluation Metrics* for Regression Model and when you would use one?

Mid

Linear Regression 48

Answer

- **Mean absolute error (MAE)**: calculates the absolute difference between actual and predicted values. It can be used when we want that our model be robust to outliers, but this metric has the disadvantage of not being differentiable so we can't use it if we want to apply optimizers like *Gradient descent*.
- **Mean squared error (MSE)**: calculates the squared difference between actual and predicted value. We can use this metric if we want to give bigger penalization to *outliers* and apply optimizers who require differentiation. MSE is a differentiable function that makes it easy to perform mathematical operations in comparison to a non-differentiable function like MAE.
- **Root mean squared error (RMSE)**: This is simply the square root of mean squared error. This metric is not so robust to outliers as the *mean absolute error* but it has the advantage to be differentiable so we can use it if we want to apply gradient descent to minimize losses.

When to use one depends on your loss function:

- **When to use MAE**: If being off by ten is just twice as bad as being off by 5. it is better to use the MAE if you don't want your performance metric to be overly sensitive to outliers.
- **When to use RMSE**: In many circumstances, it makes sense to give more weight to points further away from the mean - that is, being off by 10 is

more than twice as bad as being off by 5. In such cases, RMSE is a more appropriate measure of error.

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$

What are the *Assumptions of Linear Regression*?

Mid

Linear Regression 48

Answer

We make a few assumptions when we use linear regression to model the relationship between a response and a predictor. These assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or before we use a model to make a prediction.

There are **four principal assumptions** which justify the use of linear regression models for purposes of inference or prediction:

(i) Linearity and Additivity of the relationship between dependent and independent variables:

- (a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.

- (b) The slope of that line does not depend on the values of the other variables.
- (c) The effects of different independent variables on the expected value of the dependent variable are additive.

(ii) Statistical Independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)

(iii) Homoscedasticity (constant variance) of the errors

- (a) versus time (in the case of time series data)
- (b) versus the predictions
- (c) versus any independent variable

(iv) Normality of the error distribution.

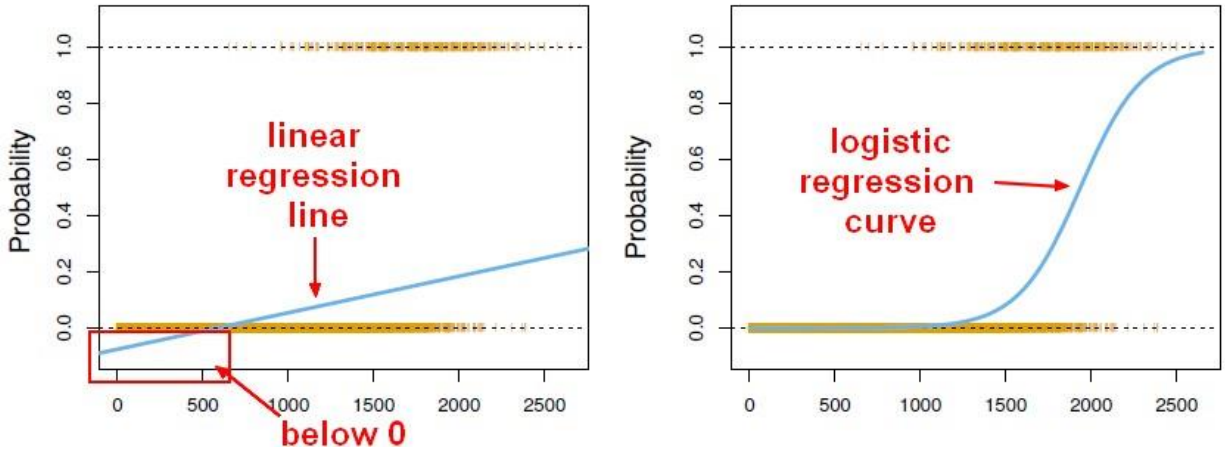
What is the difference between *Linear Regression* and *Logistic Regression*?

Mid

Linear Regression 48

Answer

- **Linear regression output as probabilities** In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values. In logistic regression, the outcome (dependent variable) has only a limited number of possible values.



- **Outcome** In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values. In logistic regression, the outcome (dependent variable) has only a limited number of possible values.
- **The dependent variable** Logistic regression is used when the response variable is categorical in nature. For instance, yes/no, true/false, red/green/blue, 1st/2nd/3rd/4th, etc. Linear regression is used when your response variable is continuous. For instance, weight, height, number of hours, etc.
- **Equation** Linear regression gives an equation which is of the form Y = mX + C, means equation with degree 1. However, logistic regression gives an equation which is of the form Y = eX + e-X
- **Coefficient interpretation** In linear regression, the coefficient interpretation of independent variables are quite straightforward (i.e. holding all other variables constant, with a unit increase in this variable, the dependent variable is expected to increase/decrease by xxx). However, in logistic regression, depends on the family (binomial, Poisson, etc.) and link (log, logit, inverse-log, etc.) you use, the interpretation is different.
- **Error minimization technique** Linear regression uses *ordinary least squares* method to minimise the errors and arrive at a best possible fit, while logistic regression uses *maximum likelihood* method to arrive at the solution. Linear regression is usually solved by minimizing the least squares error of the model to the data, therefore large errors are penalized quadratically. Logistic regression is just the opposite. Using the logistic loss function causes large errors to be penalized to an asymptotically

constant. Consider linear regression on categorical {0, 1} outcomes to see why this is a problem. If your model predicts the outcome is 38, when the truth is 1, you've lost nothing. Linear regression would try to reduce that 38, logistic wouldn't (as much)[2].

What is the difference between *Ordinary Least Squares* and *Lasso regression?*

Mid

Linear Regression 48

Answer

- **Ordinary least squares** fit a linear model by minimizing the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation. Mathematically it solves a problem of the form:
- $\min_w ||Xw - y||_2^2$
- The **Lasso** regression fits a linear model that estimates **sparse coefficients**. It is useful in some contexts due to its tendency to prefer solutions with fewer non-zero coefficients, effectively reducing the number of features upon which the given solution is dependent. Mathematically, it consists of a linear model with an added regularization term:

$$\min_w \frac{1}{2n}||Xw - y||_2^2 + \alpha||w||_1$$

Why can't a *Linear Regression* be used instead of *Logistic Regression*?

Mid

Linear Regression 48

Answer

- It is required for the independent and dependent variables to be **linear** for linear regression models, but the independent and dependent variables are **not** required to have a linear relationship in logistic functions.
- The **Linear Regression** models assume that the error terms are *normally distributed* (bell-shaped graph) whereas there are *no error terms* in **Logistic Regression** because it is assumed to follow a *Bernoulli distribution*.
- **Linear regression** has a *continuous output*. **Logistic regression** does *not* have a continuous output, rather the output is a probability between 0 and 1. A linear regression may have an output that can go beyond 0 and 1.

Why use *Root Mean Squared Error (RMSE)* instead of *Mean Absolute Error (MAE)*?

Mid

Linear Regression 48

Answer

This depends on your **loss function**. In many circumstances it makes sense to give more weight to points further away from the mean:

- If being off by 10 is more than *twice* as bad as being off by 5. In such cases, **RMSE** is a more appropriate measure of error.
- If being off by 10 is just twice as bad as being off by 5, then **MAE** is more appropriate.

Another situation when you want to use (R)MSE instead of MAE: when your observations' conditional distribution is asymmetric and you want an unbiased fit. The (R)MSE is minimized by the conditional *mean*, the MAE by the conditional *median*. So if you minimize the MAE, the fit will be closer to the median and biased.

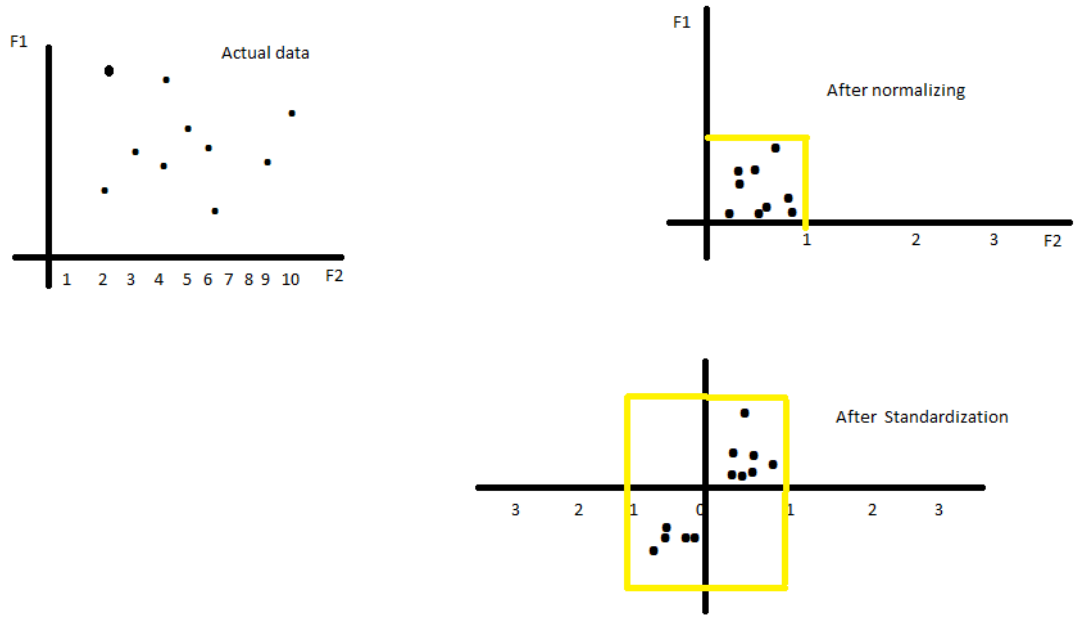
Why would you use *Normalisation* vs *Standardisation* for Linear Regression?

Mid

Linear Regression 48

Answer

- **Normalization** transforms your data into a range between 0 and 1
- **Standardization** transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1



Normalization/standardization are designed to achieve a similar goal, which is to create features that have similar ranges to each other. We want that so we can be sure we are capturing the true information in a feature and that we don't overweigh a particular feature just because its values are much larger than other features.

If all of your features are within a similar range of each other then there's no real need to standardize/normalize. If, however, some features naturally take on values that are much larger/smaller than others then normalization/standardization is called to fix it.

If you're going to be normalizing at least one variable/feature, I would do the same thing to all of the others as well.

Explain the *Stepwise Regression* technique

Senior

Linear Regression 48

Answer

Stepwise Regression is a **feature selection** technique which objective is to reduce the number of features and, hence, reduce the computational complexity of the model. This technique is based on select models with the lowest **p-values**.

For illustrated this technique let's suppose we got 6 predictors in the dataset, so in order to perform stepwise regression we must follow the next steps:

1. We fit the model with **one predictor** and the target variable. We tried each predictor one by one for then compute its p-value. Let's say that among all predictors the model with the lowest p-value was the one that contains only X1, so we keep this model.
2. Now will fit the model with **two predictors**. One we have already selected in step 1 and for the second predictor, we will try one by one with all remaining predictors. In other words, we fit one model using X1 and X2, another model using X1 and X3, and so on. For each case we compute the p-value and once again we select the model with the lowest p-value.
3. Now will try to fit the model with **three predictors**. We take the predictors already selected in step 2 and the third could it be any of the remaining ones. But if in this process we found that for each possible model we no longer reach a p-value less than 0.05 we **stop** this process. A p-value greater than 0.05 means that the model is not significant so we can reject it.

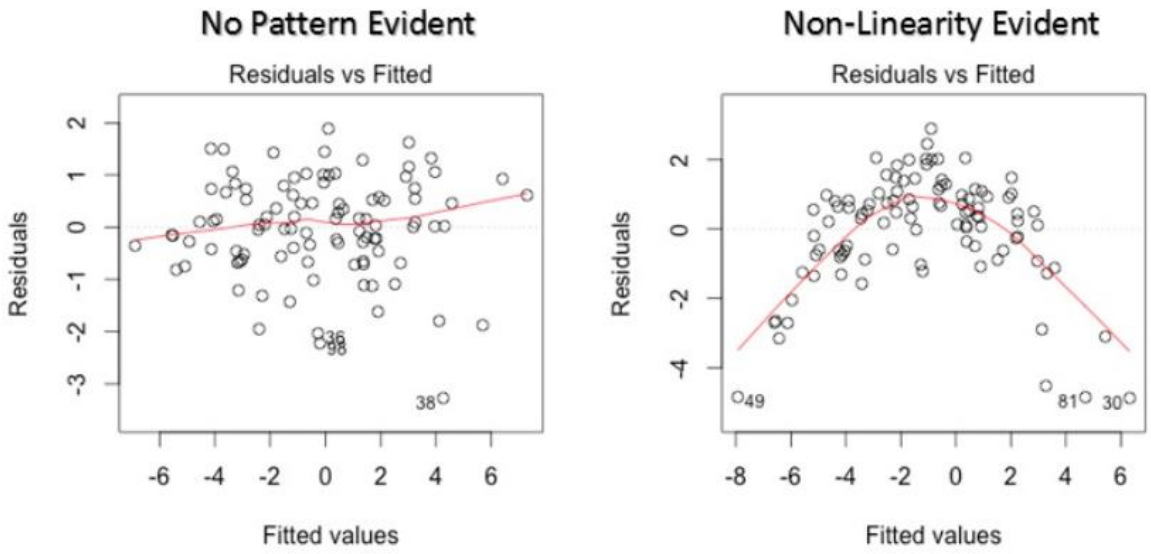
By following the previous steps we can get the **smallest set** of features that have a significant impact on the final model fit, and at the same time, reduce computational cost and avoid overfitting.

How would you check if a Linear Model follows all Regression assumptions?

Senior

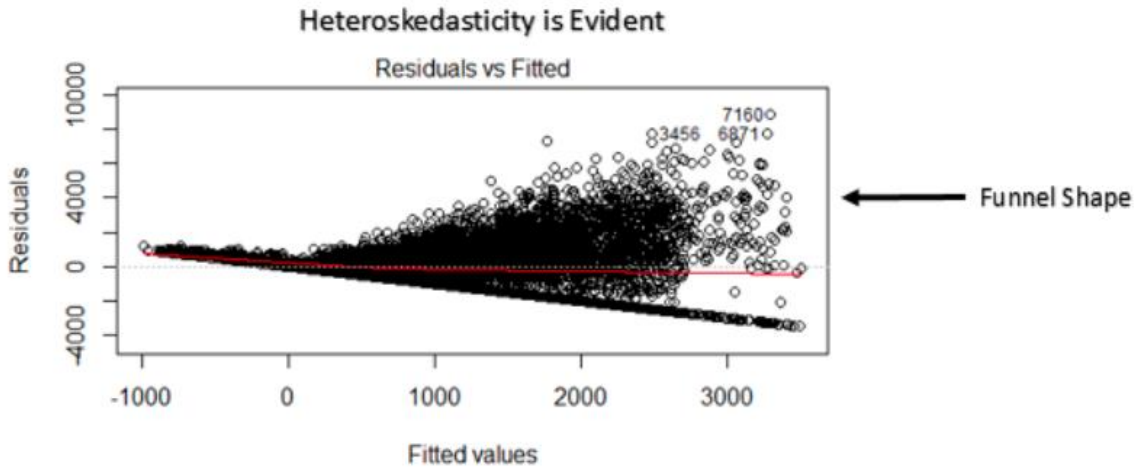
Answer

- For check **linearity and additivity**: we plot the *residual values* vs *fitted values*. If there exists any pattern (maybe, a parabolic shape) in this plot, we can consider it as signs of non-linearity in the data.

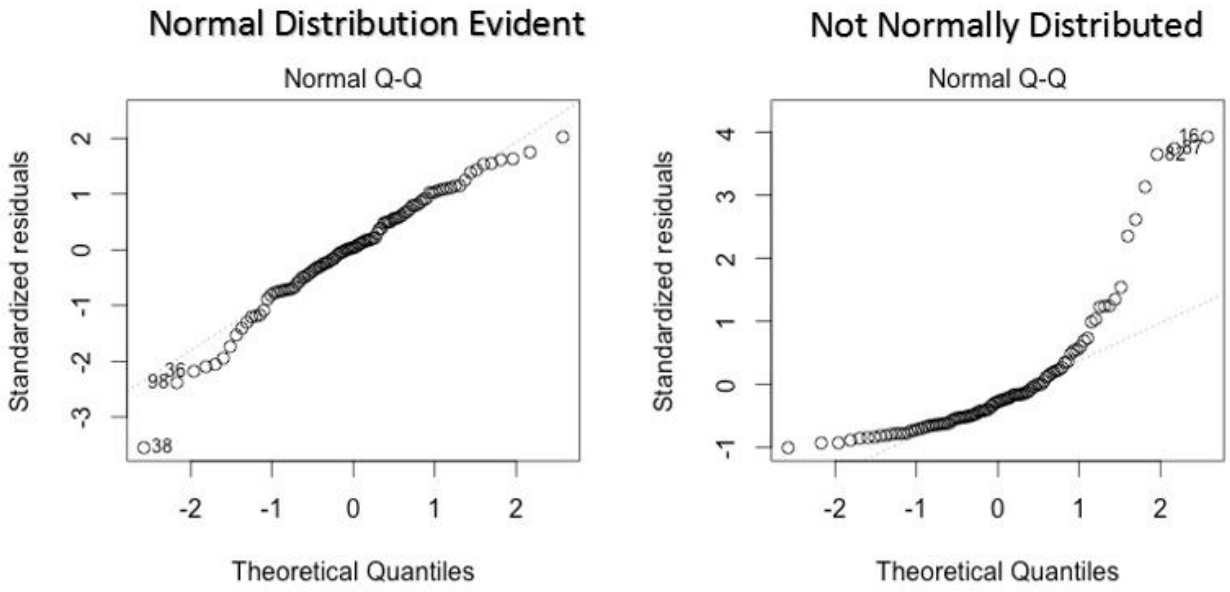


- For check **Autocorrelation**: we can calculate the *Durbin-Watson* (DW) statistic. Depending on its value we may be faced:
  - No autocorrelation if DW = 2.
  - Positive autocorrelation if 0 < DW < 2.
  - Negative autocorrelation if 2 < DW < 4.
- For check **Multicollinearity**: we can use *scatterplots* or calculate the *Variance Inflation Factor* (VIF).
  - If VIF = 1: we have complete absence of collinearity.
  - If VIF = 1 to 5: we are faced moderate correlation.
  - If VIF > 10: we have high multicollinearity.

- For check **Homoscedasticity**: we can plot the *residual values* vs the *fitted values*. If homoscedasticity is not present, the plot would exhibit a *funnel shape* pattern.



- For check **Normal Distribution of error terms**: we can use a *Q-Q plot*, if the data comes from a *normal distribution* the plot would show a fairly straight line. Otherwise, we would see a deviation in the straight line.



How would you deal with *Overfitting* in Linear Regression models?

Senior

Linear Regression 48

Answer

Overfitting is a synonym of a *complex model*, so we can solve this problem by trying to reduce its complexity. For this purpose, we can perform some **regularization techniques**, these techniques add a *penalty term* to the best fit derived from the trained data, in order to achieve a *lesser variance* with the tested data. It also restricts the influence of predictor variables over the output variable by compressing their coefficients and then reducing the complexity of the model. The regularization techniques are:

1. **Ridge Regression**: It works by adding bias to a multilinear regression model. The penalty term is known as ***L1*** and is defined as $\lambda(m)^2$ where m is the slope of the line.

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \lambda(m)^2$$

The penalty term restricts the coefficients of predictor variables but **never makes them zero**. In this way, we will have a better accurate regression with tested data at a cost of losing accuracy for the training data.

1. **Lasso Regression**: It works in a similar way that ridge regression but only differs in the penalty term ***L2***, which is equal to $\lambda|m|$. This regression is defined below as:

In this case, the penalty term can remove the variables by making their **coefficients to zero** thus removing the variables that have *high covariance* with other predictor variables.

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \lambda|m|$$

1. **ElasticNet regression**: This is a fancier combination of both Ridge and Lasso. A hyperparameter α with values between 0 and 1, is provided to assign how much weight is given to each of the ***L1*** and ***L2*** penalties:

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + \alpha\lambda(m)^2 + (1-\alpha)\lambda|m|$$

The parameter α determines the mix of the penalties and is often pre-chosen on qualitative grounds. This technique can result in better performance than a model with either one or the other penalty depending on the problem and the complexity of the model.

Does correlation imply causation? Why or why not?

No, while correlation is popularly used to provide information on the extent and direction of the linear relationship between two variables and can be used to determine whether a variable can be used to predict another, a high correlation does not imply causation.

For instance, you might find a correlation between umbrella malfunctions and a carpenter's income. As you can imagine, it is unlikely that there is a direct relation between the two, except that people tend to open up their umbrellas during the rainy season and that wooden doors swell due to high humidity. In this case, there is a hidden cause, rain, that causes both the phenomena as mentioned above and consequently the high correlation between them.

Is linear regression suitable for time series analysis?

While linear regression can be used for time series analysis and generally yield workable results, the performance is not particularly remarkable. The two main factors for this are :

- Time series generally have seasonal or periodic trends (such as peak seasons or even peak hours), which might be treated as outliers in linear regression and hence not appropriately accounted for.
- Future prediction is a generally sought-after use case in time series analysis, which will require extrapolation and rarely results in good predictions.

  Is Feature Engineering necessary for even simple linear regression? Explain with an example.
- Yes, Feature Engineering could be helpful even with the most straightforward linear regression problems. Say, for instance, you are trying to predict the Cost of a chocolate bar given the following

Length Breadth Cost

3.0	2.0	7.5
3.5	2.0	7.6
3.5	2.5	8.0
5.0	3.0	9.0

- Here you might find a workable relationship between the Length and the Cost or the Breadth and the Cost. However, on multiplying the Length and the Breadth to derive the Size, you will see that this is a much better indicator of the Cost and will fit the resulting linear regression model better.

Are there any risks to extrapolation? Explain with an example when you would and would not use this.

- Extrapolation is essentially predicting values of the target function for parameter values outside the range of those observed during training. While extrapolation could reasonably work well in some cases, such as predicting the voltage in Ohm's law, it can also result in meaningless results.
- One easy example to explain this can be to extrapolate the decreasing rainfall trend at the end of the rainy season. If the extrapolation is done unchecked, the model could predict negative rain after a period that is about as nonsensical as it gets!

Can linear regression be used for representing quadratic equations?

Yes, paradoxically, a multiple linear regression model can be used to represent quadratic equations. For more complex linear regression models, we use multiple independent variables to predict the dependent variable. Such a linear regression model is called a multiple linear regression model.

A linear model with multiple dependent variables x1, x2, ..., xn can be written as

y=1x1+2x2+...+nxn.

For a quadratic function given by, say, y=ax2+bx+c, we can use x1=x2, x2=x, and x3=1, effectively representing the desired quadratic equation. Similarly, linear regression models can be used to describe higher-order polynomials as well.

Give an example scenario where a multiple linear regression model is necessary.

Consider an example where you are considering customer satisfaction for a particular brand of cereal. This would usually be decided by several factors, including cost, nutritional value, and taste. Say you are given all the above parameters and choose x1, x2, and x3 to represent them.

If these are the only three dependent variables, then your linear regression model, in this case, would be a multiple linear regression model that can be represented in the form

$y=1x1+2x2+3x3$

Is Overfitting a possibility with linear regression?

Yes, Overfitting is possible even with linear regression. This happens when multiple linear regression is used to fit an extremely high-degree polynomial. When the parameters of such a model are learned, they will fit too closely to the training data, fitting even the noise, and thereby fail to generalize on test data.

Is it necessary to remove outliers? Why or why not?

Yes, it is necessary to remove outliers as they can have a huge impact on the model's predictions. Take, for instance, plots 3 and 4 for the Anscombe's quartet provided above. It is apparent from these plots that the outliers have caused a significant change in the best fit line in comparison to what it would have been in their absence.

How do you identify outliers?

An outlier is an observation that is unlikely to have occurred in a data set under ordinary circumstances. Its values are so widely different from the other observations that it is most likely a result of noise or a rare exceptional case.

Box plots are a simple, effective, and hence commonly used approach to identifying outliers. Besides this, scatter plots, histograms and Z-scores are other methods used whenever feasible.

How do residuals help in determining the quality of a model?

Residuals are the deviations of the observed values from a fitted line. Checking the residuals is an important step to ascertain whether our assumptions of the regression model are valid. Suppose there is no apparent pattern in the plot of

residuals versus fitted values, and the ordered residuals result in an almost normal distribution. In that case, we can conclude that there are no apparent violations of assumptions. On the other hand, if there is a relationship between the residuals and our fitted values, it is an indicator that the model is not good.

What is scaling? When is it necessary?

The technique to standardize the features in the data set to fit within a fixed range is called Feature Scaling. It is performed during the preprocessing stage and helps avoid the dominance of certain features due to high magnitudes.

When using the analytical solution for Ordinary Least Square, feature scaling is almost useless. However, when using gradient descent as an optimization technique, the data scaling results can be valuable. It can help to ensure that the gradient descent moves smoothly towards the global minimum and that the gradient descent steps update at the same rate for all the features.

If your training error is 10% and your test error is 70%, what do you infer?

A low error in training error while the test data yields a significantly higher error is a strong indicator of Overfitting. Such an observation strongly suggests that the model has learned so well over the training set that it hardly makes any mistakes during prediction over training data but cannot generalize over the unseen test set.

If you have two choices of hyperparameters, one resulting in a training and test error of 10% and another with a training and test error of 20%, which one of the two would you prefer and why?

Given that both the training and the test set are yielding an error of 10% in case 1 and an error of 20% in case 2, it is pretty easy to opt for the hyperparameters of case 1 for our machine learning problem as it is always desirable to have a lower error in predictions.

If your training error is high despite adjusting the hyperparameter values and increasing the number of iterations, what is most likely to be the issue? How can you resolve this problem?

High training error despite hyperparameter adjustment and a significant number of iterations strongly indicates that the model is unable to learn the problem it is presented with despite its best effort, or in other words, that it is underfitting.

Reducing the regularisation and using more complex models can be some ways used to address this problem.

If the deviations of the residuals from your model are extremely small, does it suggest that the model is good or bad?

Residuals are essentially how much the actual data points vary from the fitted line and are hence indicators of deviation or error. Therefore, the smaller the deviation of the residuals from the fitted line, the better the model is likely to be.

What scenario would you prefer to use Gradient Descent instead of Ordinary Least Square Regression and why?

Ordinary Least Square Regression is computationally very expensive. Therefore, while it performs well with small data sets, it is infeasible to use this approach for significant machine learning problems. Consequently, for problems with larger data sets, Gradient Descent is the preferred optimization algorithm.

If you observe that the test error is increasing after a certain number of iterations, what do you infer is most likely to be occurring? How do you address this problem?

Observing an increase in error on the validation set after a certain number of iterations can indicate that the model is Overfitting. We can arrive at this diagnosis because we expect the error to decrease with more optimized parameters. While simplifying the model is one way to address this problem, early stopping is another commonly used solution.

Early stopping is probably one of the most commonly used forms of regularization. Unlike a weight decay used in the cost function, which helps to arrive at less complex models by explicit regularization, early stopping can be considered as a form of implicit regularization.

What are the important assumptions of Linear regression?

A linear relationship

Restricted Multi-collinearity value

Homoscedasticity

Firstly, there has to be a linear relationship between the dependent and the independent variables. To check this relationship, a scatter plot proves to be useful.

Secondly, there must no or very little multi-collinearity between the independent variables in the dataset. The value needs to be restricted, which depends on the domain requirement.

The third is the homoscedasticity. It is one of the most important assumptions which states that the errors are equally distributed.

2. What is heteroscedasticity?

Heteroscedasticity is exactly the opposite of homoscedasticity, which means that the error terms are not equally distributed. To correct this phenomenon, usually, a log function is used.

What are the possible ways of improving the accuracy of a linear regression model?

There could be multiple ways of improving the accuracy of a linear regression, most commonly used ways are as follows:

1. Outlier Treatment:

-Regression is sensitive to outliers, hence it becomes very important to treat the outliers with appropriate values. Replacing the values with mean, median, mode or percentile depending on the distribution can prove to be useful.

How to interpret a Q-Q plot in a Linear regression model?

A Q-Q plot is used to check the normality of errors. In the above chart mentioned, Majority of the data follows a normal distribution with tails curled. This shows that the errors are mostly normally distributed but some observations may be due to significantly higher/lower values are affecting the normality of errors.

What is the significance of an F-test in a linear model?

– The use of F-test is to test the goodness of the model. When the model is re-iterated to improve the accuracy with changes, the F-test values prove to be useful in terms of understanding the effect of overall regression.

What are the disadvantages of the linear model?

– Linear regression is sensitive to outliers which may affect the result.

– Over-fitting

– Under-fitting

What are the basic assumptions of the Linear Regression Algorithm?
The basic assumptions of the Linear regression algorithm are as follows:

- **Linearity:** The relationship between the features and target.
- **Homoscedasticity:** The error term has a constant variance.
- **Multicollinearity:** There is no multicollinearity between the features.
- **Independence:** Observations are independent of each other.
- **Normality:** The error(residuals) follows a normal distribution.

Now, let's break these assumptions into different categories:

Assumptions about the form of the model:

It is assumed that there exists a linear relationship between the dependent and the independent variables. Sometimes, this assumption is known as the **'linearity assumption**'.

Assumptions about the residuals:

- **Normality assumption:** The error terms, $\varepsilon(i)$, are normally distributed.
- **Zero mean assumption:** The residuals have a mean value of zero.
- **Constant variance assumption:** The residual terms have the same (but unknown) value of variance, σ^2. This assumption is also called the assumption of homogeneity or homoscedasticity.
- **Independent error assumption:** The residual terms are independent of each other, i.e. their pair-wise covariance value is zero.

Assumptions about the estimators:

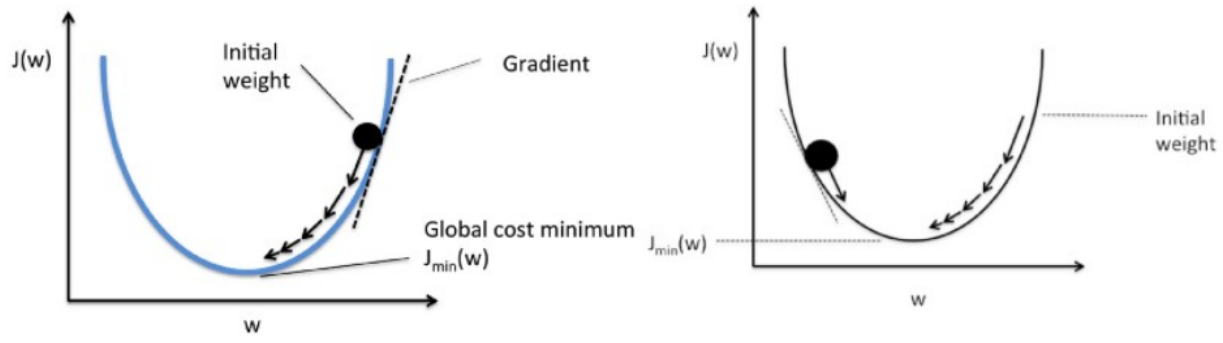
- The independent variables are measured without error.
- There does not exist a linear dependency between the independent variables, i.e. there is no multicollinearity in the data.

4. Explain the difference between Correlation and Regression.
Correlation: It measures the strength or degree of relationship between two variables. It doesn't capture causality. It is visualized by a single point.

Regression: It measures how one variable affects another variable. Regression is all about model fitting. It tries to capture the causality and describes the cause and the effect. It is visualized by a regression line.

Explain the Gradient Descent algorithm with respect to linear regression.
Gradient descent is a **first-order optimization algorithm**. In linear regression, this algorithm is used to optimize the cost function to find the values of the β_s **(estimators)** corresponding to the optimized value of the cost function.

The working of Gradient descent is similar to a **ball that rolls down a graph (ignoring the inertia)**. In that case, the ball moves along the direction of the maximum gradient and comes to rest at the flat surface i.e, corresponds to minima.



Justify the cases where the linear regression algorithm is suitable for a given dataset.

Generally, a Scatter plot is used to see if linear regression is suitable for any given data. So, we can go for a linear model if the relationship looks somewhat linear. Plotting the scatter plots is easy in the case of simple or univariate linear regression.

But if we have more than one independent variable i.e, the case of multivariate linear regression, then two-dimensional pairwise scatter plots, rotating plots, and dynamic graphs can be plotted to find the suitableness.

On the contrary, to make the relationship linear we have to apply some transformations.

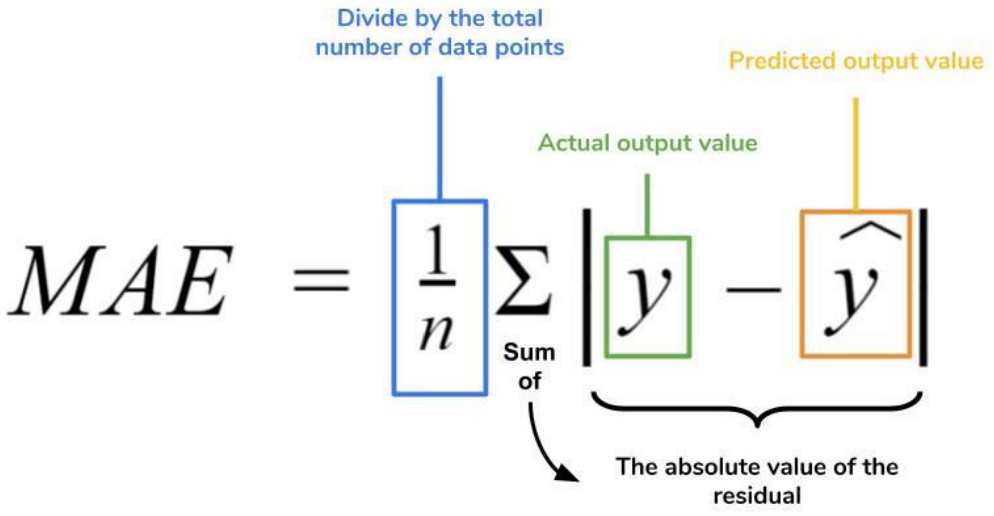
List down some of the metrics used to evaluate a Regression Model.
Mainly, there are five metrics that are commonly used to evaluate the regression models:

- Mean Absolute Error(MAE)
- Mean Squared Error(MSE)
- Root Mean Squared Error(RMSE)
- R-Squared(Coefficient of Determination)
- Adjusted R-Squared

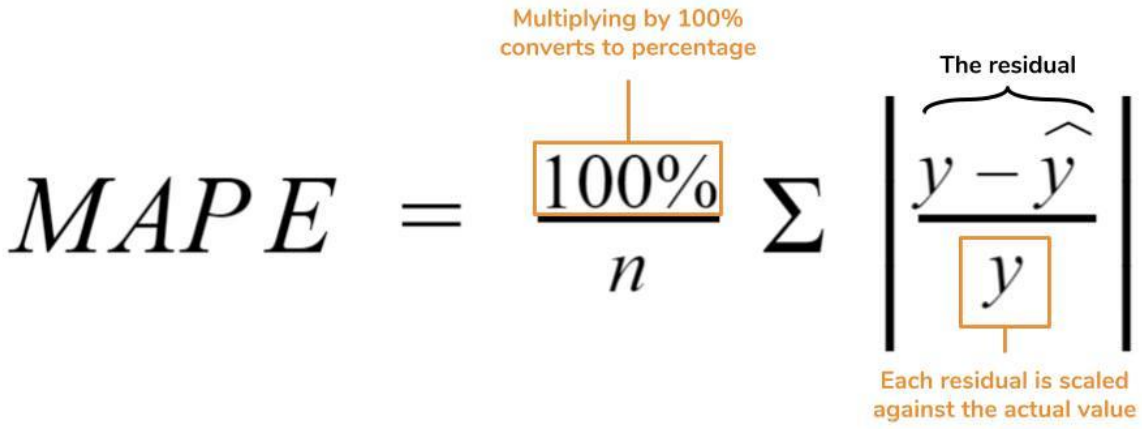
What is OLS?
OLS stands for **Ordinary Least Squares**. The main objective of the linear regression algorithm is to find coefficients or estimates by minimizing the error term i.e, **the sum of squared errors**. This process is known as OLS.

This method finds the best fit line, known as regression line by minimizing the sum of square differences between the observed and predicted values.

What are MAE and MAPE?
MAE stands for **Mean Absolute Error**, which is defined as the average of absolute or positive errors of all values. In simple words, we can say MAE is an average of absolute or positive differences between predicted values and the actual values.



MAPE stands for **Mean Absolute Percent Error**, which calculates the average absolute error in percentage terms. In simple words, It can be understood as the percentage average of absolute or positive errors.



Why do we square the residuals instead of using modulus?
This question can be understood that **why one should prefer the absolute error instead of the squared error.**

1. In fact, the absolute error is not differentiable, hence cannot used in gradient descent optimization.

2. Moreover in mathematical terms, the squared function is differentiable everywhere, while the absolute error is not differentiable at all the points in its domain(its derivative is undefined at 0). This makes the squared error more preferable to the techniques of mathematical optimization. To optimize the squared error, we can compute the derivative and set its expression equal to 0, and solve. But to optimize the absolute error, we require more complex techniques having more computations.

List down the techniques that are adopted to find the parameters of the linear regression line which best fits the model.
There are mainly two methods used for linear regression:

1. Ordinary Least Squares(Statistics domain):

To implement this in Scikit-learn we have to use the **LinearRegression()** class.

2. Gradient Descent(Calculus family):

To implement this in Scikit-learn we have to use the **SGDRegressor()** class.

Which evaluation metric should you prefer to use for a dataset having a lot of outliers in it?
Mean Absolute Error(MAE) is preferred when we have too many outliers present in the dataset because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and these start penalizing the outliers by squaring the error terms, commonly known as residuals.

Explain the normal form equation of the linear regression.
The normal equation for linear regression is :

$$\beta=(X^TX)^{-1}X^TY$$

This is also known as the **closed-form solution** for a linear regression model.

where,

$Y=\beta^TX$ is the equation that represents the model for the linear regression,

Y is the dependent variable or target column,

β is the vector of the estimates of the regression coefficient, which is arrived at using the normal equation,

X is the feature matrix that contains all the features in the form of columns. The thing to note down here is that the first column in the X matrix consists of all 1s, to incorporate the offset value for the regression line.

17. When should it be preferred to the Gradient Descent method instead of the Normal Equation in Linear Regression Algorithm?
To answer the given question, let's first understand the difference between the Normal equation and Gradient descent method for linear regression:

<u>Gradient descent:</u>

- Needs hyper-parameter tuning for alpha (learning parameter).
- It is an iterative process.
- Time complexity- $O(kn^2)$
- Preferred when n is extremely large.

Normal Equation:

- No such need for any hyperparameter.
- It is a non-iterative process.
- Time complexity- $O(n^3)$ due to evaluation of X^TX.
- Becomes quite slow for large values of n.

where,

'k' represents the maximum number of iterations used for the gradient descent algorithm, and

'n' is the total number of observations present in the training dataset.

Clearly, if we have large training data, a normal equation is not preferred for use due to very high time complexity but for small values of 'n', the normal equation is faster than gradient descent.

What are R-squared and Adjusted R-squared?
R-square (R^2), also known as the **coefficient of determination** measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model.

$$Coefficient\ of\ Determination \rightarrow \quad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The main problem with the R-squared is that it will always remain the same or increases as we are adding more independent variables. Therefore, to overcome this problem, an Adjusted-R^2 square comes into the picture by penalizing those adding independent variables that do not improve your existing model.

What are the flaws in R-squared?
There are two major flaws of R-squared:

Problem- 1: As we are adding more and more predictors, R^2 always increases irrespective of the impact of the predictor on the model. As R^2 always increases and never decreases, it can always appear to be a better fit with the more independent variables(predictors) we add to the model. This can be completely misleading.

Problem- 2: Similarly, if our model has too many independent variables and too many high-order polynomials, we can also face the problem of over-fitting the data. Whenever the data is over-fitted, it can lead to a misleadingly high R^2 value which eventually can lead to misleading predictions.

What is Multicollinearity?
It is a phenomenon where two or more independent variables(predictors) are highly correlated with each other i.e. one variable can be linearly predicted with the help of other variables. It determines the inter-correlations and inter-association among independent variables. Sometimes, multicollinearity can also be known as collinearity.

Reasons for Multicollinearity:

- Inaccurate use of dummy variables.
- Due to a variable that can be computed from the other variable in the dataset.

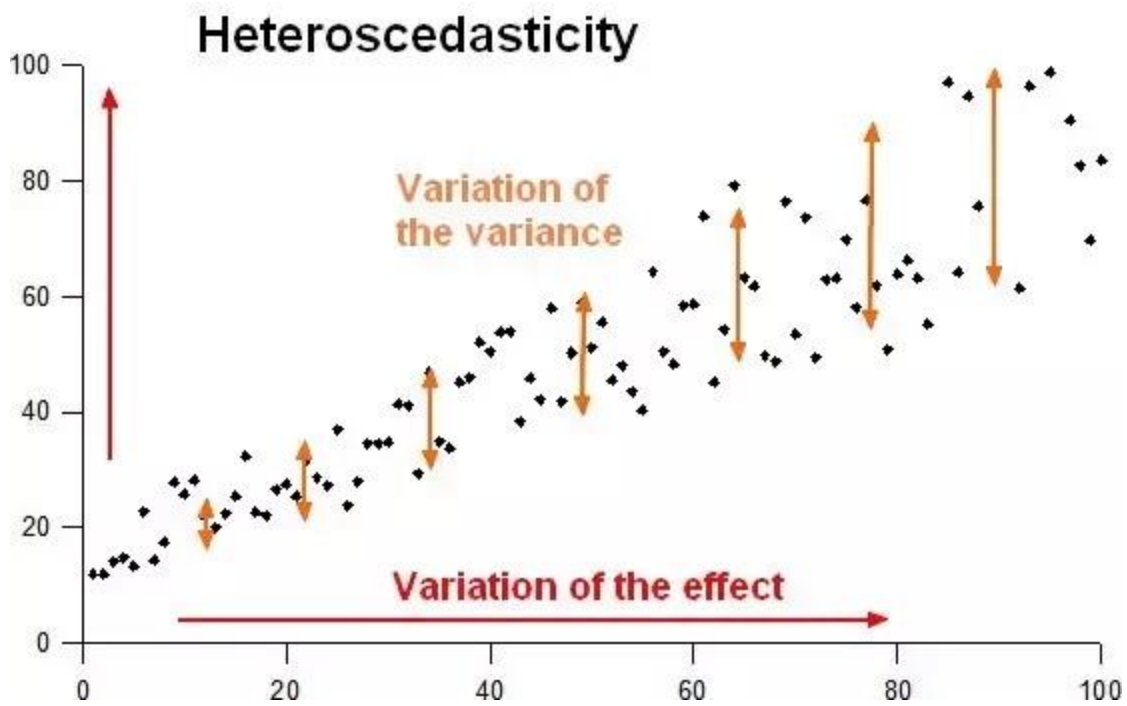
Impacts of Multicollinearity:

- Impacts regression coefficients i.e, coefficients become indeterminate.
- Causes high standard errors.

Detecting Multicollinearity:

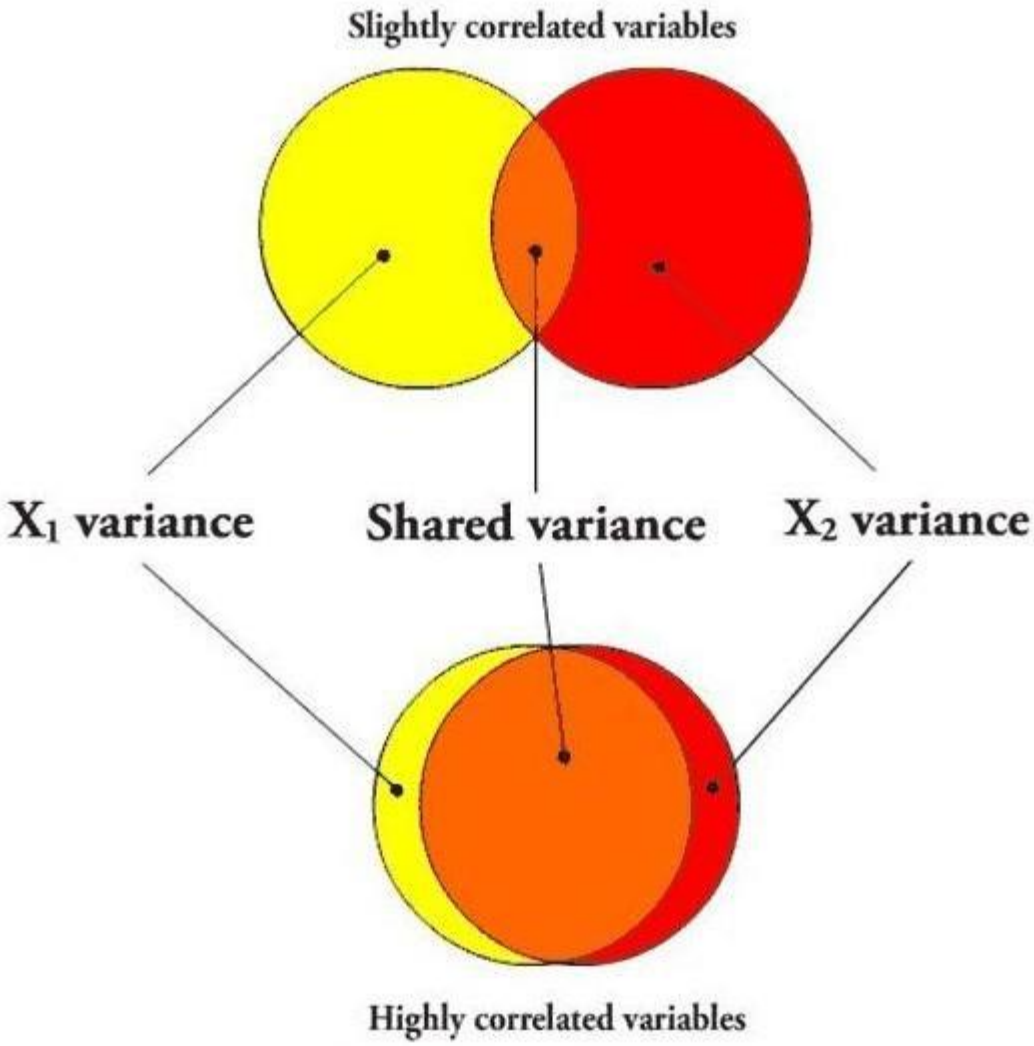
- By using the correlation coefficient.
- With the help of Variance inflation factor (VIF), and Eigenvalues.

What is Heteroscedasticity? How to detect it?
It refers to the situation where the variations in a particular independent variable are unequal across the range of values of a second variable that tries to predict it.



What are the disadvantages of the linear regression Algorithm?
The main disadvantages of linear regression are as follows:

- **Assumption of linearity:** It assumes that there exists a linear relationship between the independent variables(input) and dependent variables (output), therefore we are not able to fit the complex problems with the help of a linear regression algorithm.
- **Outliers:** It is sensitive to noise and outliers.
- **Multicollinearity:** It gets affected by multicollinearity.



To learn more about, R^2 and adjusted-R^2, refer to the

What is VIF? How do you calculate it?
VIF stands for **Variance inflation factor**, which measures how much variance of an estimated regression coefficient is increased due to the presence of collinearity between the variables. It also determines how much multicollinearity exists in a particular regression model.

How is Hypothesis testing used in Linear Regression Algorithm?
For the following purposes, we can carry out the Hypothesis testing in linear regression:

1. To check whether an independent variable (predictor) is significant or not for the prediction of the target variable. Two common methods for this are —

By the use of p-values:

If the p-value of a particular independent variable is greater than a certain threshold (usually 0.05), then that independent variable is insignificant for the prediction of the target variable.

---

1. **Consider equations with the fewest number of predictor/explanatory variables if models that are being compared are nearly equivalent in terms of significance and fit**
    - a) Law of MLR
    - b) Law of equations
    - c) Law of Parsimony
    - d) Law of Parsinomy

    Answer - c) Law of Parsimony
1. **In MLR, if there are three variable the shape is**
    - a) Circle
    - b) Hyperplane
    - c) Triangle
    - d) None of the above

    Answer - b) Hyperplane
1. **If there are more than three variables, it is not possible to visualize them.**
    - a) True
    - b) False

    Answer - a) True
1. **In MLR the shape is not really a line.**
    - a) True
    - b) False

    Answer - a) True
1. **To visualize all the pair of variables at a time, we use**

- a) Scatter plot b) Box plot c) Scatter plot matrix d) None of the above

Answer - c) Scatter plot matrix

1. **Reducing the number of features and computational complexity of the model is done by**

   - a) Feature selection
   - b) Cross selection
   - c) Complete selection
   - d) None of the above

   Answer - a) Feature selection

1. **In regression _____ are the feature selection methods used .**

   - a) Forward selection
   - b) Backward elimination
   - c) Stepwise regression
   - d) All the above

   Answer - d) All the above

1. **We use the _____ plot to know whether there is a relationship between two variables.**

   - a) Histogram
   - b) Box plot
   - c) Scatter plot
   - d) Bar plot

   Answer - c) Scatter plot

1. **After fitting the model with all the required variables_____ method is used for identifying and removing independent variables that do not contribute enough to the model.**

   - a) Forward selection
   - b) Backward elimination
   - c) Stepwise regression
   - d) All the above

   Answer - b) Backward elimination

1. **Considering variables one by one and building the model by checking the significance value & R square is done by using _____ method.**

   - a) Stepwise Regression
   - b) Backward elimination
   - c) Both b and c
   - d) Stepwise elimination

   Answer - a) Stepwise Regression

1. **Using the regression model, we can estimate the strength and direction of the association from the adjusted partial regression of Independent Variables**

    - a) True
    - b) False

    Answer - a) True

_____increases only when independent variable is significant and affects dependent variable.

- a) R-squared
- b) Adjusted R-squared
- c) Both the above
- d) None of the above

Answer - b) Adjusted R-squared

1. **A lower AIC or BIC value indicates a _____ fit.**

    - a) Worst fit
    - b) Better fit
    - c) Low fit
    - d) None of the above

    Answer - b) Better fit

1. **Will you be able to improve your linear regression model by making it more complex i.e. by adding more linear regression variables to it?**

    - a) Yes
    - b) No

    Answer - b) No

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. It's generally used where the target variable is Binary or Dichotomous.

What is a logistic function? What is the range of values of a logistic function?

$f(z) = 1/(1+e^{-z})$

The values of a logistic function will range from 0 to 1. The values of Z will vary from -infinity to +infinity.

Why is logistic regression very popular?

Logistic regression is famous because it can convert the values of logits (logodds), which can range from -infinity to +infinity to a range between 0 and 1. As logistic functions output the probability of occurrence of an event, it can be applied to

many real-life scenarios. It is for this reason that the logistic regression model is very popular.

What are odds?

It is the ratio of the probability of an event occurring to the probability of the event not occurring. For example, let's assume that the probability of winning a lottery is 0.01. Then, the probability of not winning is 1- 0.01 = 0.99.

The odds of winning the lottery = (Probability of winning)/(probability of not winning)

The odds of winning the lottery = 0.01/0.99

The odds of winning the lottery is 1 to 99, and the odds of not winning the lottery is 99 to 1.

What are the outputs of the logistic model and the logistic function?

The logistic model outputs the logits, i.e. log odds; and the logistic function outputs the probabilities.

Logistic model = α+1X1+2X2+….+kXk. The output of the same will be logits.

Logistic function = $f(z) = 1/(1+e^{-(\alpha+1X1+2X2+....+kXk)})$. The output, in this case, will be the probabilities.

$$\ln\left(\frac{P(Y)}{1\text{-}P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

$$\Updownarrow$$

$$P(Y) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K\right)}$$

- Assumptions in logistic regression
  - Y_i are from Bernoulli or binomial (n_i, μ_i) distribution

    Y_i are independent

The $\beta 0$ and $\beta 1$ values are estimated during the training stage using *maximum-likelihood* estimation or *gradient descent*. Once we have it, we can make predictions by simply putting numbers into the *logistic regression equation* and calculating a result.

When *Logistic Regression* can be used?

Junior

Ⓔ Logistic Regression 24

Answer

Logistic regression can be used in *classification* problems where the output or dependent variable is *categorical* or *binary*. However, in order to implement logistic regression correctly, the dataset must also satisfy the following properties:

1. There should not be a high correlation between the independent variables. In other words, the predictor variables should be independent of each other.
2. There should be a linear relationship between the logit of the outcome and each predictor variable. The logit function is given as $\text{logit}(p) = \log(p/(1-p))$, where p is the probability of the outcome.
3. The sample size must be large. How large depends on the number of independent variables of the model.

When all the requirements above are satisfied, logistic regression can be used.

Why is Logistic Regression called *Regression* and not *Classification*?
Junior

Ⓔ Logistic Regression 24

Answer

Although the task we are targeting in logistic regression is a classification, *logistic regression does not actually individually classify things for you*: it just gives you probabilities (or log odds ratios in the logit form).

The only way logistic regression can actually classify stuff is if you apply a rule to the probability output. For example, you may round probabilities greater than or equal to 50% to 1, and probabilities less than 50% to 0, and that's your classification.

Compare *SVM* and *Logistic Regression* in handling outliers
Mid

⟶ SVM 56

Answer

- For **Logistic Regression**, outliers can have an *unusually large effect* on the estimate of logistic regression coefficients. It will find a linear boundary if it exists to accommodate the outliers. To solve the problem of outliers, sometimes a sigmoid function is used in logistic regression.
- For **SVM**, outliers can make the decision boundary deviate severely from the optimal hyperplane. One way for SVM to get around the problem is to intrduce *slack variables*. There is a penalty involved with using slack variables, and how SVM handles outliers depends on how this penalty is imposed.

How a *Logistic Regression* model is trained?

Mid

Ⓔ Logistic Regression 24

Answer

The **logistic model** is trained through the **logistic function**, defined as:

$$P(y) = \frac{1}{1 + e^{-wx}}$$

where x is the input data, w is the weight vector, y is the output label, and P(y) is the probability that the output label belongs to one class. If for some input we got P(y) > 0.5, then the predicted output is 1, and otherwise would be 0.
The training is based in estimate the w vector. For this, in each training instance we use ***Stochastic Gradient Descent*** to calculate a prediction using some initial values of the coefficients, and then calculate new coefficient values based on the error in the previous prediction. The process is repeated for a fixed number of iterations or until the model is accurate enough or cannot be made any more accurate.

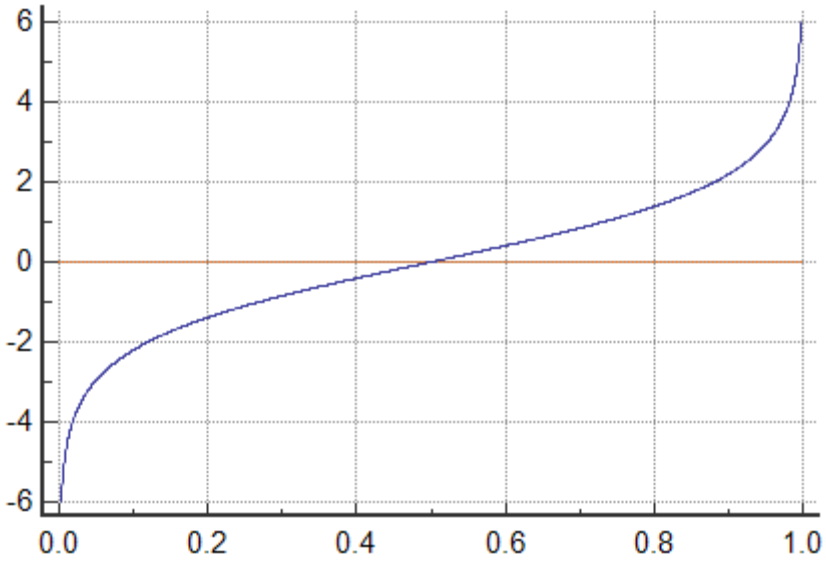
How do you use a supervised *Logistic Regression* for Classification?

Mid

Ⓔ Logistic Regression 24

Answer

- ***Logistic regression*** is a statistical model that utilizes ***logit*** function to model classification problems. It is a regression analysis to conduct when the dependent variable is *binary*. The **logit** function is shown below:



- Looking at the logit function, the next question that comes to mind is *how to fit that graph/equation*. The fitting of the logistic regression is done using the *maximum likelihood* function.
- In a supervised logistic regression, **features** are mapped onto the **output**. The output is usually a categorical value (which means that it is mapped with one-hot vectors or binary numbers).
- Since the **logit** function always outputs a value between 0 and 1, it gives the **probability of the outcome**.

Provide a mathematical intuition for Logistic Regression?

Mid

Ⓔ Logistic Regression 24

Answer

Logistic regression can be seen as a **transformation** from linear regression to logistic regression using the logistic function, also known as the **sigmoid function** or S(x):

$$S(x) = \frac{1}{1 + e^{-x}}$$

Given the linear model:

$$y = b_0 + b_1 \cdot x$$

If we apply the sigmoid function to the above equation it results:

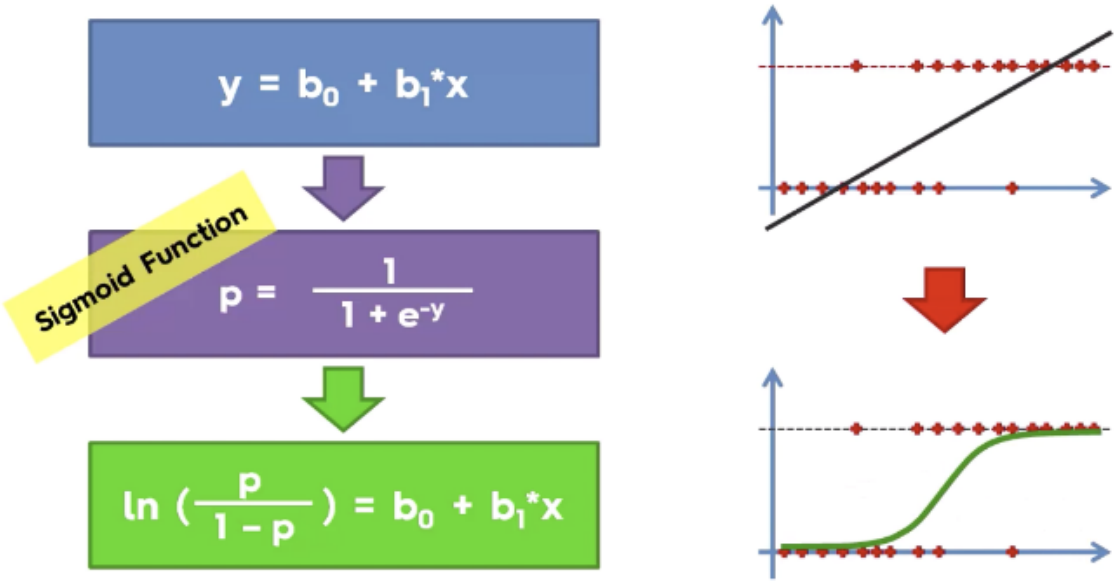
$$S(y) = \frac{1}{1+e^{-y}} = p$$

where p is the probability and it takes values between 0 and 1. If we now apply the logit function to p, it results:

$$logit(p) = log(\frac{p}{1-p}) = b_0 + b_1 \cdot x$$

The equation above represents the logistic regression. It fits a logistic curve to set of data where the dependent variable can only take the values 0 and 1.

The previous transformation can be illustrated in the following figure:



logistic function (also called the 'inverse logit').

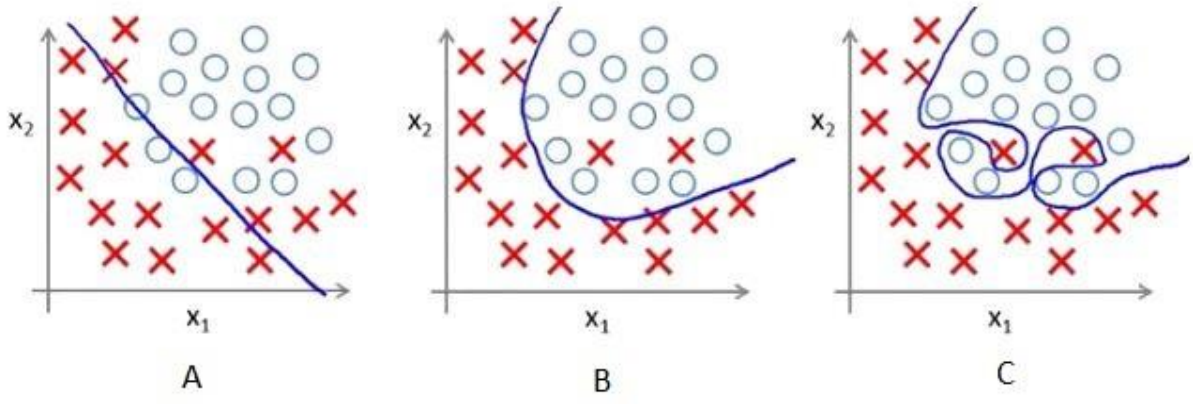
What can you infer from each of the hand drawn decision boundary of *Logistic Regression* below?

Mid

Ⓔ Logistic Regression 24

Problem

Also, what should we do to fix the problem of each decision boundary?



Answer

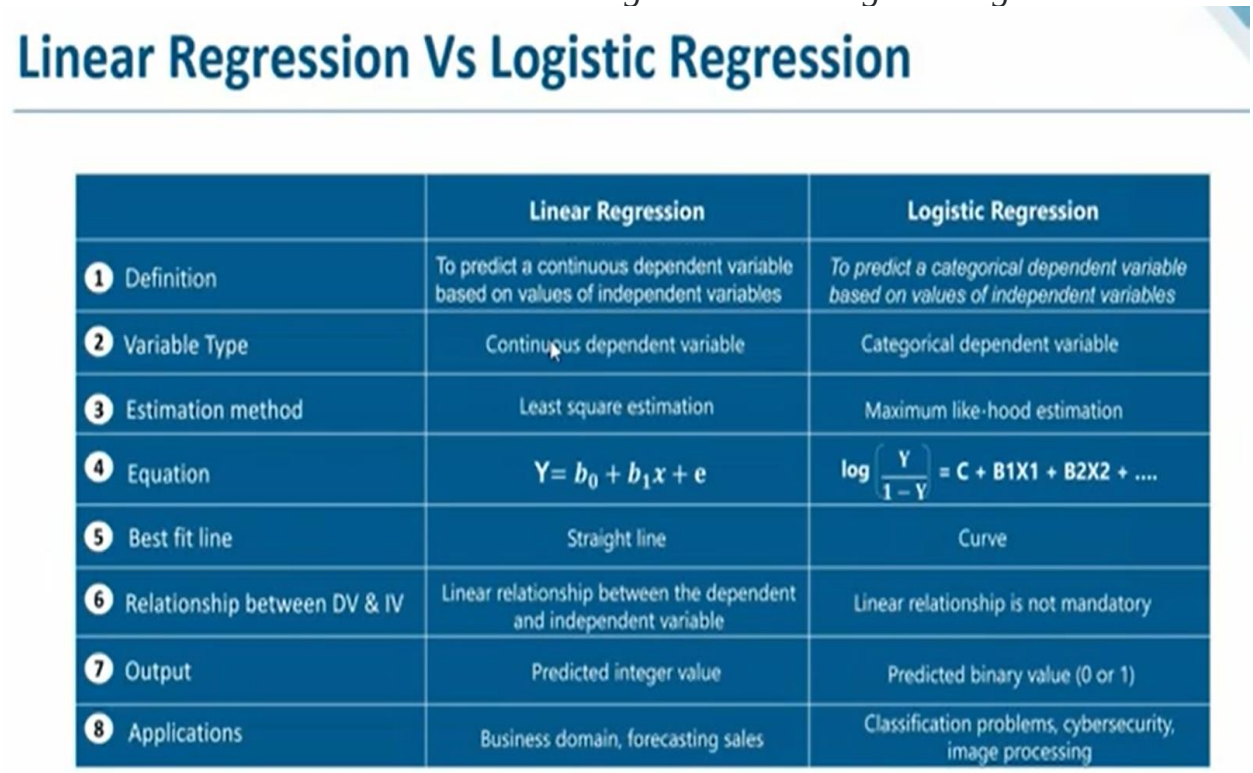
What can we infer:

- **A**: the model *underfits* the data. It will give us the maximum error compared to other two models.
- **B**: *best-fitting* model.
- **C**: the model *overfits* the data. It performs exceptionally well on training data but performs considerably worse on test data.

What can we do to fix the problem:

- **A**: *increase* the complexity of the model or *increase* the number of independent variables.
- **B**: best performing model, so we don't need to tweak anything.
- **C**: add *regularization* method to the model.

What is the difference between *Linear Regression* and *Logistic Regression*?



Error minimization technique Linear regression uses *ordinary least squares* method to minimise the errors and arrive at a best possible fit, while logistic regression uses *maximum likelihood* method to arrive at the solution.

- **Softmax function**:
    - Is used for *multi-class* classification in logistic regression models, when we have only *one right answer* or *mutually exclusive* outputs.
    - Its probabilities sum will be 1.
    - Is used in *different layers* of neural networks.
    - It is defined as:

$$\mathrm{softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \forall \mathrm{j} = 1 \ldots \mathrm{K}$$

- **Sigmoid function**:
    - Is used for *multi-label* classification in logistic regression models, when we have *more than one right answer* or *non-exclusive* outputs.
    - Its probabilities sum does not need to be 1.

- Is used as an *activation function* while building neural networks.
- It is defined as:

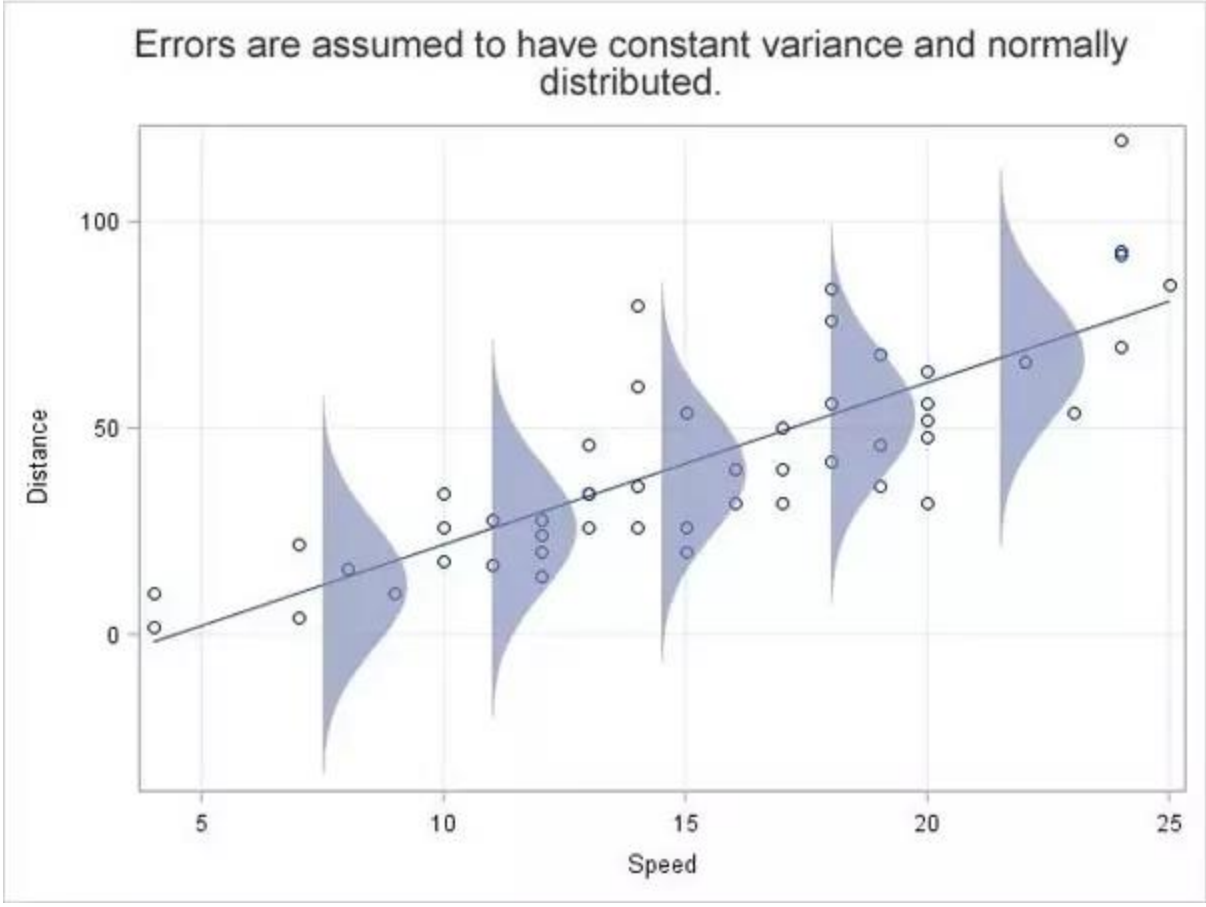
$$\sigma(z_j) = \frac{e^{z_j}}{1 + e^{z_j}}$$

Why can't a *Linear Regression* be used instead of *Logistic Regression*?
Mid

Linear Regression 48

Answer

- It is required for the independent and dependent variables to be **linear** for linear regression models, but the independent and dependent variables are **not** required to have a linear relationship in logistic functions.
- The **Linear Regression** models assume that the error terms are *normally distributed* (bell-shaped graph) whereas there are *no error terms* in **Logistic Regression** because it is assumed to follow a *Bernoulli distribution*.



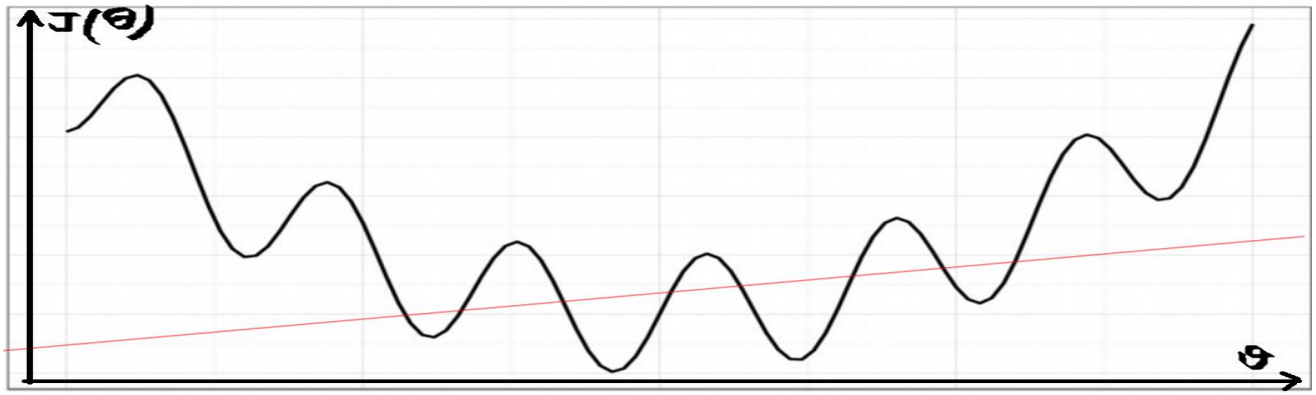
Why don’t we use *Mean Squared Error* as a cost function in Logistic Regression?

Mid

Ⓔ Logistic Regression 24

Answer

- In Linear Regression, we used the **Squared Error** mechanism.
- For **Logistic Regression**, such a cost function produces a **non-convex** space with many local minimums, in which it would be very difficult to minimize the cost value and find the global minimum.



Why is *Logistic Regression* considered a *Linear Model*?

Mid

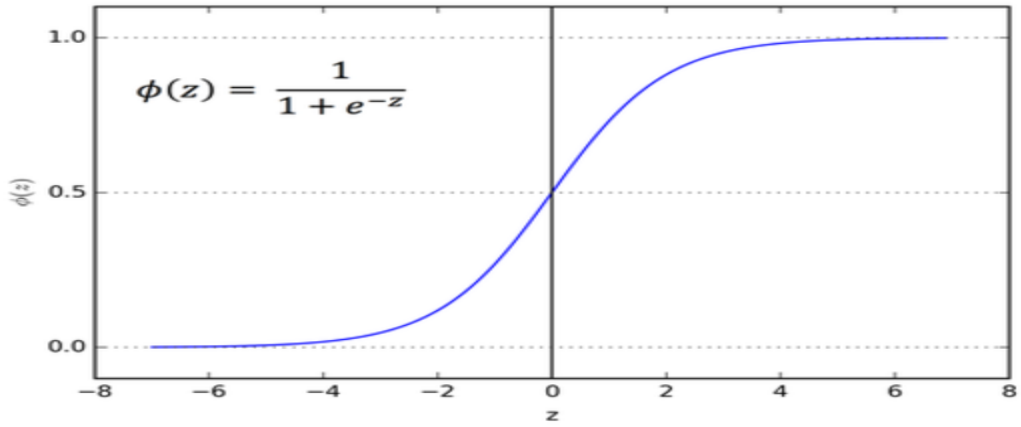
Ⓔ Logistic Regression 24

Answer

A model is considered linear if the **transformation of features** that is used to calculate the prediction is a **linear combination of the features**. Although Logistic Regression uses **Sigmoid function** which is a nonlinear function, the model is a generalized linear model because the outcome always depends on the sum of the **inputs and parameters**.

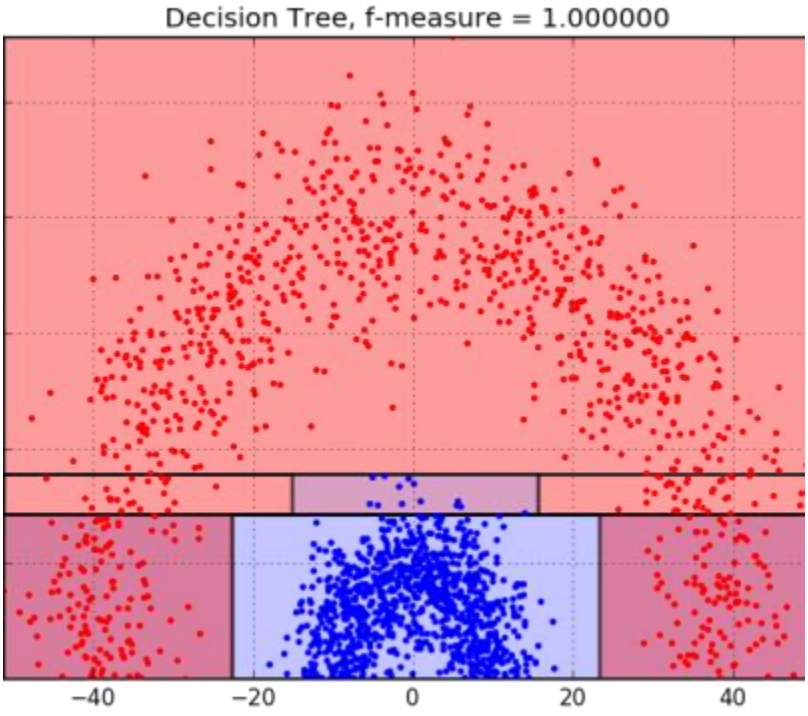
i.e the logit of the estimated probability response is a linear function of the predictors parameters.

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i}$$

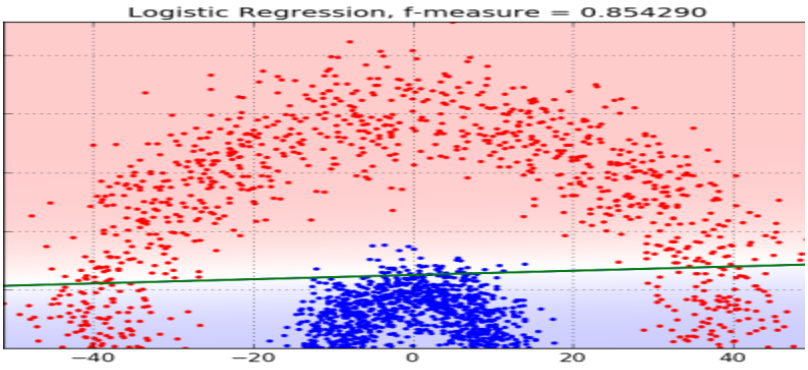


Compare *Decision Trees* and *Logistic Regression*
ecision Boundaries

- ***Decision trees*** bisect the output space into smaller and smaller regions. Even though trees fit the data better, it is prone to overfitting.



- ***Logistic regression*** fits a single line to divide the space exactly into two. In higher dimensions, this line would generalize to planes and hyperplanes.



Compare *Naive Bayes* vs with *Logistic Regression* to solve classification problems Naïve bayes is suitable for smaller dataset and there is conditional independence among the features.

Naive Bayes assumes conditional independence among the features, which is not true in most real-life problems. So then, if for a problem these assumptions are not satisfied, the Logistic Regression model would be less biased and therefore, it will outperform Naive Bayes when given lots of training data.

Naive Bayes has a time complexity of O(log n), whereas the logistic regression is O(n). So, for n number of features, Naive Bayes converges more quickly to its asymptotic estimates. Hence, it will outperform Logistic Regression in the case of a small training dataset.

How can we avoid *Over-fitting* in *Logistic Regression* models?

Senior

Ⓔ Logistic Regression 24

Answer

Regularization techniques can be used to avoid over-fitting in regression models.

Two types of regularization techniques used for logistic regression models are *Lasso Regularization, and Ridge Regularization*. Ridge and Lasso allow the regularization (*shrinking*) of coefficients. This reduces the *variance* in the model

which contributes to the model not overfitting on the data. Ridge and Lasso add penalty values to the loss function as shown below:

1. **Lasso regression** adds the absolute value of the magnitude of coefficient as penalty term to the loss function as can be seen in the equation below:

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} |w_i|$$

The second term added to the loss function is a penalty term whose size depends on the *total magnitude of all the coefficients*. Lambda is a *tuning parameter* that adjusts how large a penalty there will be.

1. **Ridge** regression adds *squared magnitude* of all the coefficients as penalty term to the loss function as can be seen in the equation below:

$$Loss = Error(y, \hat{y}) + \lambda \sum_{i=1}^{N} w_i^2$$

The extra penalty term in this case disincentivizes including extra features. There is a balancing act between the increasing of a coefficient and the corresponding increase to the overall variance of the model.

The Ridge and Lasso act as their own feature selector were features that don't drive the predictive power of the regression to see their coefficients pushed down, while more predictive features see higher coefficients despite the added penalty.

Imagine that you know there are *outliers* in your data, would you use *Logistic Regression*?

Senior

••• Anomaly Detection 47

Answer

- **Logistic Regression:** Logistic Regression is an algorithm that will be highly *influenced* by outliers. If we have outliers in our dataset, the decision boundary could be shifted, which will lead to incorrect inferences. So using Logistic Regression wouldn't be optimal if we have outliers.
- **Tree-based model:** Algorithms like **decision trees** or **random forests** are more **robust to outliers**. This is because unlike a linear model like Logistic Regression, a tree-based model works by splitting the data into groups (repeatedly) according to whether a data point is above or below a selected threshold value on a selected feature variable. Thus, a data point with a much higher value than the rest wouldn't influence the decision of a tree-based model at all.

Name some advantages of using *Support Vector Machines* vs *Logistic Regression* for classification

Senior

⟶ **SVM** 56

Answer

The goal of ***Support vector machine (SVM)*** and ***Logistic regression (LR)*** is to find a hyperplane that distinctly classifies the data points. The main difference between these two algorithms is in their *loss functions*: *SVM* minimizes ***hinge loss*** while *logistic regression* minimizes ***logistic loss***.

- **Hinge-Loss** - When used for Standard SVM, the loss function denotes the size of the margin between linear separator and its closest points in either class. Only differentiable everywhere with p = 2

$$\max\left[1 - h_{\mathbf{w}}(\mathbf{x}_i)y_i, 0\right]^p$$

- **Log-loss** - One of the most popular loss functions in Machine Learning, since its outputs are well-calibrated probabilities.

$$\log(1 + e^{-h_{\mathbf{w}}(\mathbf{x}_i)y_i})$$

Things to remember:

- The *logistic loss* diverges faster than *hinge loss*. So, in general, *SVM* will be more robust to outliers than *LR*.
- The *logistic loss* does not go to zero even if the point is classified sufficiently confidently, so this might lead to minor degradation in accuracy. *Hinger loss* can take the value of 0, so with *SVM* we don't have this problem.
- *LR* produces probabilistic values while *SVM* produces 1 or 0. So in a few words, *LR* makes no absolute prediction and it does not assume data is enough to give a final decision.
- *SVM* tries to maximize the margin between the closest support vectors while *LR* maximizes a class probability. Thus, *SVM* finds a solution that is as *fair as possible* for the two categories while *LR* has not this property.
- *SVM* use kernel tricks and transform datasets into rich features that can classify no *linearly separable* data while *LR* does not perform well on this kind of problems.

When would you use *SVM* vs *Logistic regression*?

Senior

⟶ **SVM** 56

Answer

Generally, the first approach is to try with ***logistic regression*** to see how the model performs. If we found that the data can't be linearly separable or the model does not have a good performance, then we can try using ***SVM***.

We can also considerer the following rule of thumb: Given n number of *features* and m number of *training* examples:

- If n is less than 10,000 and m is less than 1,000 we can use either *logistic regression* or *SVM* with a *linear kernel*.
- If n is less than 1,000 and m is less than 10,000 we can use *SVM* with a *Gaussian*, *polynomial*, or another kernel.

- If n is less than 1,000, and m is greater than 50,000 we first add more features manually and then can use *logistic regression* or *SVM* with a *linear kernel*.

How do you implement multinomial logistic regression?

- The multinomial logistic classifier can be implemented using a generalization of the sigmoid, called the softmax function. The softmax represents each class with a value in the range (0,1), with all the values summing to 1. Alternatively, you could use the one-vs-all or one-vs-one approach using multiple simple binary classifiers.

Suppose that you are trying to predict whether a consumer will recommend a particular brand of chocolate or not. Let us say your hypothesis function outputs h(x)=0.55 where h(x) is the probability that y=1 (or that a consumer recommends the chocolate) given any input x. Does this mean that the consumer will recommend the chocolate?

- The answer to this question is 'cannot be determined.' And this will remain the case unless you are provided additional data on the decision boundary. Let us say that you set the decision boundary such that y=1 is h(x)≥0.5 and 0; otherwise, then the answer for this question would be a resounding YES. However, if you set the decision boundary (although this is not very common practice) such that y=1 is h(x)≥0.6 and 0, otherwise the answer will be a NO.

Why can't we use the mean square error cost function used in linear regression for logistic regression?

- If we use mean square error in logistic regression, the resultant cost function will be non-convex, i.e., a function with many local minima, owing to the presence of the sigmoid function in h(x). As a result, an attempt to find the parameters using gradient descent may fail to optimize cost function properly. It may end up choosing a local minima instead of the actual global minima.

6) If you observe that the cost function decreases rapidly before increasing or stagnating at a specific high value, what could you infer?

- A trend pattern of the cost curve exhibiting a rapid decrease before then increasing or stagnating at a specific high value indicates that the learning

rate is too high. The gradient descent is bouncing around the global minimum but missing it owing to the larger than necessary step size.

Are there alternatives to find optimum parameters for logistic regression besides using Gradient Descent?

Yes, Gradient Descent is merely one of the many available optimization algorithms. Other advanced optimization algorithms can often help arrive at the optimum parameters faster and help with scaling for significant machine learning problems. A few such algorithms are Conjugate Gradient, BFGS, and L-BFGS algorithms.

How many binary classifiers would you need to implement one-vs-all for three classes? How does it work?

You would need three binary classifiers to implement one-vs-all for three classes since the number of binary classifiers is precisely equal to the number of classes with this approach. If you have three classes given by y=1, y=2, and y=3, then the three classifiers in the one-vs-all approach would consist of **h(1)(x),** which classifies the test cases as 1 or not 1, **h(2)(x)** which classifies the test cases as 2 or not 2 and so on. You can then take the results together to arrive at the correct classification. For example, with three categories, Cats, Dogs, and Rabbits, to implement the one-vs-all approach, we need to make the following comparisons:

Binary Classification Problem 1: Cats vs. Dogs, Rabbits (or not Cats)

Binary Classification Problem 2: Dogs vs. Cats, Rabbits (or not Dogs)

Binary Classification Problem 3: Rabbits vs. Cats, Dogs (or not Rabbits)

What is the importance of regularization?

Regularisation is a technique that can help alleviate the problem of overfitting a model. It is beneficial when a large number of parameters are present, which help predict the target function. In these circumstances, it is difficult to select which features to keep manually.

Regularisation essentially involves adding coefficient terms to the cost function so that the terms are penalized and are small in magnitude. This helps, in turn, to preserve the overall trends in the data while not letting the model become too complex. These penalties, in effect, restrict the influence a predictor variable can have over the target by compressing the coefficients, thereby preventing overfitting.

Why is the Wald Test useful in logistic regression but not in linear regression?

The Wald test, also known as the Wald Chi-Squared Test, is a method to find whether the independent variables in a model are of significance. The significance of variables is decided by whether they contribute to the predictions or not. The variables that add no value to the model can therefore be deleted without risking severe adverse effects to the model. The Wald test is unnecessary in linear regression because it is easy to compare a more complicated model to a simpler model to check the influence of the added independent variables. After all, we can use the R2 value to make this comparison. However, this is not possible with logistic regression as we use Maximum Likelihood Estimate, which uses the previously mentioned method infeasible. The Wald test can be used for many different models, including those with binary variables or continuous variables, and has the added advantage that it only requires estimating one model.

Will the decision boundary be linear or non-linear in logistic regression models? Explain with an example.

The decision boundary is essentially a line or a plane that demarcates the boundary between the classes to which linear regression classifies the dependent variables. The shape of the decision boundary will depend entirely on the logistic regression model.

For logistic regression model given by hypothesis function h(x)=g(Tx)where g is the sigmoid function, if the hypothesis function is h(x)=g(1+2x2+3x3)then the decision boundary is linear. Alternatively, if h(x)=g(1+2x22+3x32)then the decision boundary is non-linear.

What are odds? Why is it used in logistic regression?

Odds are the ratio of the probability of success to the probability of failure. The odds serve to provide the constant effect a particular predictor or independent variable has on the output prediction. Expressing the effect of a predictor on the

likelihood of the target having a particular value through probability does not describe this constant effect. In linear regression models, we often want to measure the unique effect of each independent variable on the output for which the odds are very useful.

Given fair die, what are the odds of occurrence of odd numbers?

The odds of occurrence of odd numbers is 1.

There are three odd and three even numbers in a fair die, and therefore, the probability of occurrence of odd numbers is 3/6 or 0.5. Similarly, the odds of occurrence of numbers that are not odd is 0.5. Since odds is the ratio of the probability of success and that of failure,

Odds = 0.5/0.5=1.

In classification problems like logistic regression, classification accuracy alone is not considered a good measure. Why?

Classification accuracy considers both true positives and false positives with equal significance. If this were just another machine learning problem of not too much consequence, this would be acceptable. However, when the problems involve deciding whether to consider a candidate for life-saving treatment, false positives might not be as bad as false negatives. The opposite can also be true in some cases. Therefore, while there is no single best way to evaluate a classifier, accuracy alone may not serve as a good measure.

It is common practice that when the number of features or independent variables is larger in comparison to the training set, it is common to use logistic regression or support vector machine (SVM) with a linear kernel. What is the reasoning behind this?

It is common to use logistic regression or SVM with a linear kernel because when there are many features with a limited number of training examples, a linear function should be able to perform reasonably well. Besides, there is not enough training data to allow for the training of more complex functions.

Between SVM and logistic regression, which algorithm is most likely to work better in the presence of outliers? Why?

SVM is capable of handling outliers better than logistic regression. SVM is affected only by the points closest to the decision boundary. Logistic regression, on the other hand, tries to maximize the conditional likelihood of the training data and is therefore strongly affected by the presence of outliers.

Which is the most preferred algorithm for variable selection?

Lasso is the most preferred for variable selection because it performs regression analysis using a shrinkage parameter where the data is shrunk to a point, and variable selection is made by forcing the coefficients of not so significant variables to be set to zero through a penalty.

What according to you is the method to best fit the data in logistic regression?

Maximum Likelihood Estimation to obtain the model coefficients which relate to the predictors and target.

Answer using either TRUE or FALSE. Is logistic regression a type of a supervised machine learning algorithm?
Ans. Yes, the answer to this question would be TRUE because, indeed, logistic regression is a supervised machine learning algorithm.
Answer using either TRUE or FALSE. Is logistic regression mainly used for classification?
Ans. Yes, the answer to this question is TRUE. Indeed, logistic regression is primarily used for classification tasks rather than performing actual regression.
Answer this question using TRUE or FALSE. Can a neural network be implemented, which mimics the behavior of a logistic regression algorithm?
Ans. Yes, the answer would be TRUE. Neural networks are also known as universal approximators. They can be used to mimic almost any machine learning algorithm.
Answer this question using either TRUE or FALSE. Can we use logistic regression to solve a multi-class classification problem?
Ans. The short answer would be TRUE. The long answer, however, would have you thinking a little. There is no way in which you can implement a multi-class classification from just using one single logistic regression model. You will need to either use a neural network with a softmax activation function or use a complex machine-learning algorithm to predict many classes of your input variable successfully.
Choose one of the options from the list below. What is the underlying method which is used to fit the training data in the algorithm of logistic regression?

1. Jaccard Distance
2. Maximum Likelihood
3. Least Square error
4. None of the options which are mentioned above.

Ans. The answer is 2. It is easy to select option C, which is the Least Square error because this is the same method that is used in linear regression.

Choose one of the options from the list below. Which metric would we not be able to use to measure the correctness of a logistic regression model?

1. The area under the receiver operating characteristics curve (or AUC-ROC score)
2. Log-loss
3. Mean squared error (or MSE)
4. Accuracy

Ans. The correct option you should choose is 3, i.e., Mean Squared Error, or MSE. Since the logistic regression algorithm is actually a classification algorithm rather than a basic regression algorithm, we cannot use the Means Square Error to determine the performance of the logistic regression model that we wrote.

Choose one of the options from the list below. AIC happens to be an excellent metric to judge the performance of the logistic regression model. AIC is very similar to the R-squared method that is used to determine the performance of a linear regression algorithm. What is actually true about this AIC?

1. The model with a low AIC score is generally preferred.
2. The model which has s huge AIC score is actually preferred.
3. The choice of the model just from the basis of the AIC score highly depends on the situation.
4. None of the options which are mentioned above.

Ans. The model which has the least value of AIC is preferred. So, the answer to the question would be option A. The main reason why we choose the model with the lowest possible value of AIC is because the penalty, which is added to regulate the performance of the model, actually does not encourage the fit to be over.

Answer using either TRUE or FALSE. Do we need to standardize the values present in the feature columns before we feed the data into a training logistic regression model?

Ans. No, we do not need to standardize the values present in the feature space, which we have to use to train the logistic regression model. So, the answer to this question would be FALSE. We choose to standardize all our values to help the function (usually gradient descent), which is responsible for making the algorithm converge on a value. Since this algorithm is relatively simple, it does not need the

amounts to be scaled for it actually to have a significant difference in its performance.

Choose one of the options from the list below. Which is the technique we use to perform the task of variable selection?

1. Ridge Regression
2. LASSO regression
3. None of the options which are mentioned
4. Both LASSO and Ridge Regression

Ans. The answer to this question is B. LASSO regression. The reason is simple, the l2 penalty, which is incurred in the LASSO regression function, has the ability to make the coefficient of some features to be zero. Since the coefficient is zero, meaning they will not have any effect in the final outcome of the function.

Choose the correct answer from the options below. The logit function is defined as the log of the odds function. What do you think the range of this logit function be in the domain of [0,1]?

1. (-infinity, +infinity)
2. (0, +infinity)
3. (-infinity, 0)
4. (0, 1)

Ans. The probability function takes the value which it is passed with and turns it into a probability. Meaning the range of any function is clamped in between zero and one. However, the odds function does one thing it takes the value from the probability function and makes the range of it from zero to infinity.
So, the effective input to the log function would be from zero to infinity.

Choose the option which you think is TRUE from the list below:

1. The error values in the case of Linear regression have to follow a normal distribution, but in the case of logistic regression, the values do not have to follow a standard normal distribution.
2. The error values in the case of Logistic regression have to follow a normal distribution, but in the case of Linear regression, the values do not have to follow a standard normal distribution.
3. The error values in the case of both Linear regression and Logistic regression has to follow a normal distribution.
4. The error values in the case of both Linear regression and Logistic regression do not have to follow a normal distribution.

Ans. The only truthful statement in the bunch of these statements is the first one. So, the answer to the question becomes the option A.

Choose the correct option(S) from the list of options down below. So, let us say that you have applied the logistic regression model into any given data. The accuracy results that you got are X for the training set and Y for the test set.

Now, you would like to add more data points to your model. So, what, according to you, should happen?

1. The Accuracy X, which we got in the training data, should increase.
2. The Accuracy X, which we got from the training data, should decrease.
3. The Accuracy Y, which we got from the test data, should decrease.
4. The accuracy Y, which we got from the test data, should increase or remain the same.

Ans. The training accuracy highly depends on the fit the model has on the data, which it has already seen and learned. So, suppose we increase the number of features fed into the model, the training accuracy X increases. In that case, the training accuracy will grow because the model will have to become more complicated to fit the data with an increased number of features properly. Whereas the testing accuracy only will increase if the feature which is added into the model is an excellent and significant feature or else the model's accuracy while testing will more or less remain the same. So, the answer to this question would be both options A and D.

Choose the right option from the following option regarding the method of one vs. all in terms of logistic regression.

1. We would need a total of n models to classify between n number of classes correctly.
2. We would need an n-1 number of models to classify between n number of classes.
3. We would need only one single model to classify between n number of classes successfully.
4. None of the options which are mentioned above.

Ans. To classify between n different classes, we are going to need n models in a One vs. All approach.

---

1. **In Multinomial Regression, the term multi refers to _____**
    - a) Fixed number of outcomes.
    - b) Binary outcome.
    - c) More than one outcome.
    - d) More than two outcomes.

    Answer - d) More than two outcomes
1. **In Multinomial Regression, the term nominal refers to _____**
    - a) They are ordinal variables.
    - b) They are nominal variables.
    - c) They are normal variables.
    - d) None of the above.

Answer - b) They are nominal variables

1. **In Multinomial Regression, they are nominal outcome variables means _____**

    - a) There is no order in the outcome.
    - b) There is order in outcome variable.
    - c) There is only two outcome variables.
    - d) There is only two outcome variables.

Answer - a) There is no order in the outcome

1. **The outcome variable is polytomous/ multiclass/ polychotomous logistic/ softmax regression/ multinomial logit/ maximum entropy classifier/ conditional maximum entropy model all of these mean -**

    - a) Only one outcome variable.
    - b) Only two outcome variables.
    - c) More than two outcome variables.
    - d) None of the above.

Answer - c) More than two outcome variables

1. **At the center of the multinomial regression analysis is the task estimating the log odds of each category. If for example, k=n categories as the reference categories, the multinomial regression estimates _____ regression functions.**

    - a) n.
    - b) n-1.
    - c) n-2.
    - d) n-3.

Answer - b) n-1

1. **Multinomial logistic regression is often considered an attractive analysis because; it does not assume normality, linearity, or homoscedasticity.**

    - a) True.
    - b) False.
    - c) It assumes only normality.
    - d) None of the above.

Answer - a) True

1. **Multinomial regression is enhanced to _____regression.**

    - a) Simple linear.
    - b) Multi linear.
    - c) Logistic.
    - d) None of the above.

Answer - c) Logistic

1. **In multinomial regression mathematical intuition, the equation of intercepts are calculated for different _____ with one considered baseline level.**

    - a) Logit model.
    - b) Decision model.
    - c) Linear model.
    - d) None of the above.

    Answer - a) Logit model

1. **In Python ,when we are working with problems with more than two classes, you should specify the multi_class parameter of _____.**

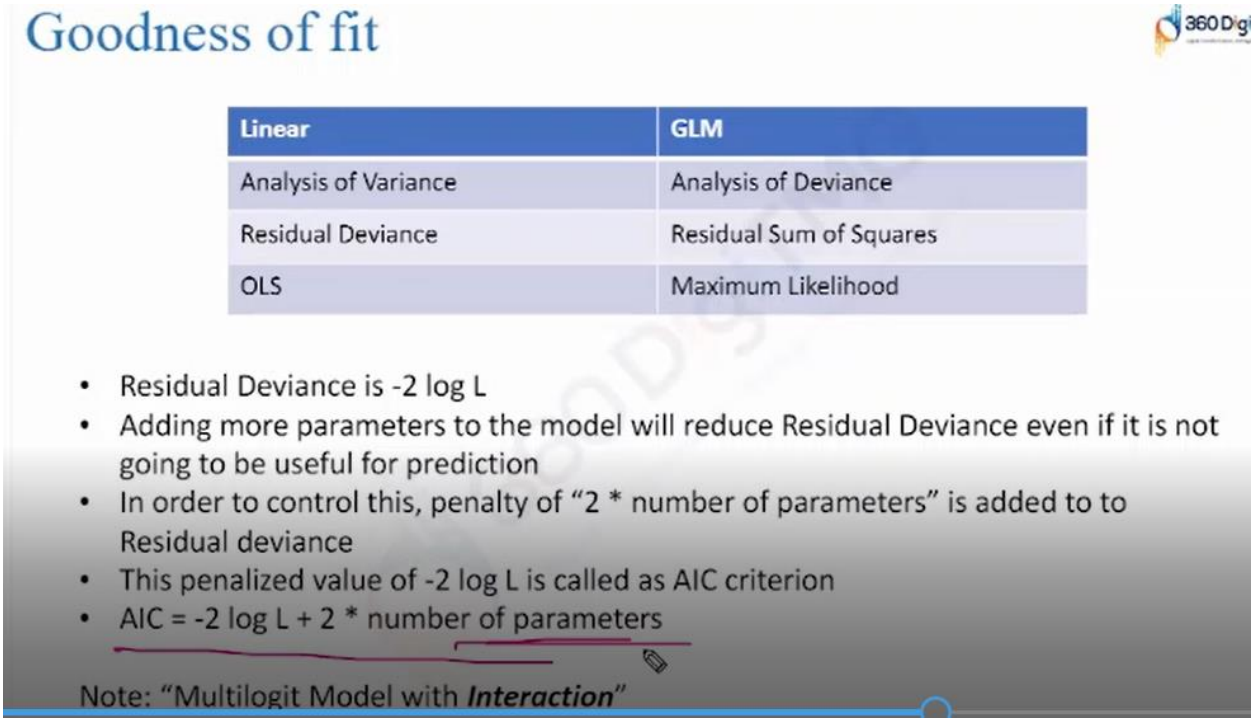
    - a) LogisticRegression.
    - b) LinearRegression.
    - c) Multilinear Regression.
    - d) None of the above.

    Answer - a) LogisticRegression

1. **In Python, "Multinomial" class option is supported only by the _____solvers.**

    - a) Saga and liblinear.
    - b) Sag and lbfgs.
    - c) Newton-cg and sag.
    - d) Lbfgs and newton-cg.

    Answer - d) Lbfgs and newton-cg



1. **In Python, for example -**

   "model = LogisticRegression(solver='liblinear',c=0.05, multi_class='ovr',random_state=0)" ● 'ovr' says to make_____.
   - a) Without any fit for each class.
   - b) The binary fit for each class.
   - c) With only a single model fit.
   - d) None of the above.

   Answer - b) The binary fit for each class

1. **In multinomial regression choose correct statement-**

   I. In multinomial logistic regression, we use the concept of one vs rest classification using binary classification technique of logistic regression.
   II. Now, for example, let us have "K" classes. First, we divide the classes into two parts, "1 "represents the 1st class and "0" represents the rest of the classes, then we apply binary classification in this 2 class and determine the probability of the object to belong in 1st class vs rest of the classes.
   III. we apply this technique for the "k" number of classes and return the class with the highest probability. By, this way we determine in which class the object belongs. In this way multinomial logistic regression works

- a) Only statement (I) is correct.
- b) Only statement (II) is correct.
- c) Only statement (II) is correct.
- d) All of the above statements are correct.

Answer - d) All of the above statements are correct

1. **While doing multinomial regression, choose the list of things which we must check to ensure that the final output is valid from below statements –**

   I. Your dependent variable must be Nominal. This does not mean that multinomial regression cannot be used for the ordinal variable. However, for multinomial regression, we need to run ordinal logistic regression.
   II. You must convert your categorical independent variables to dummy variables.
   III. There should be no multicollinearity.
   IV. There should be a linear relationship between the dependent variable and continuous independent variables. As we cannot measure this directly between nominal and continuous variables what we do is we take logit transformation of the dependent variable.
   V. Ensure that we do not have outliers and high influential points in the data.

   - a) Only statement (I).
   - b) Only statement (II).
   - c) Only statement (III).
   - d) All of the above 5 statements.

   Answer - d) All of the above 5 statements

There should be no Outliers in the data points.

Assumptions of multinomial regression:

When you want to choose multinomial logistic regression as the classification algorithm for your problem, then you need to make sure that the data should satisfy some of the assumptions required for multinomial logistic regression.

1. **The Dependent variable should be either nominal or ordinal variable.**

Nominal variable is a variable that has two or more categories but it does not have any meaningful ordering in them. For example, (a) 3 types of cuisine i.e.

Indian, Continental and Italian. (b) 5 categories of transport i.e. Bus, Car, Train, Ship and Airplane.

Ordinal variable are variables that also can have two or more categories but they can be ordered or ranked among themselves. For example, Grades in an exam i.e. A-excellent, B-Good, C-Needs Improvement and D-Fail. When ordinal dependent variable is present, one can think of ordinal logistic regression.

- Set of one or more Independent variables can be **continuous, ordinal or nominal.**

Continuous variables are numeric variables that can have infinite number of values within the specified range values. For example, age of a person, number of hours students study, income of an person.

Ordinal variables should be treated as either continuous or nominal.

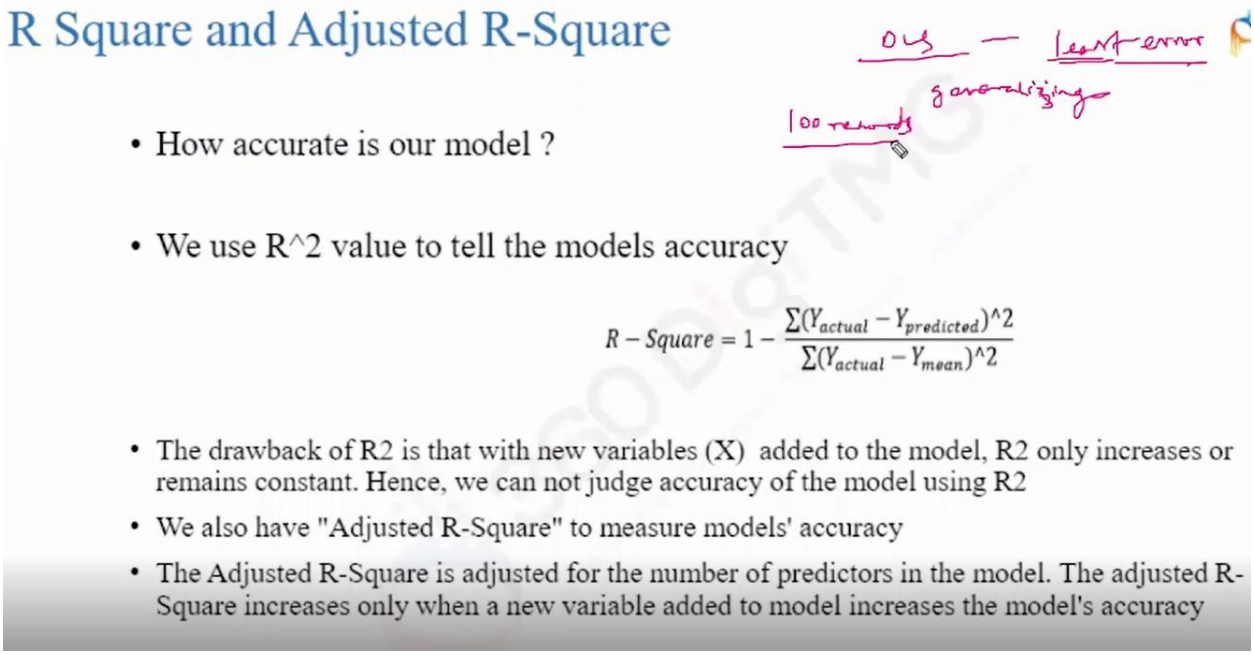
- The Observations and dependent variables must be mutually exclusive and exhaustive.

Mutually exclusive means when there are two or more categories, no observation falls into more than one category of dependent variable.

The categories are exhaustive means that every observation must fall into some category of dependent variable.

- **No Multicollinearity between Independent variables.**

Multicollinearity occurs when two or more independent variables are highly correlated with each other. This makes it difficult to understand how much every independent variable contributes to the category of dependent variable. Also makes it difficult to understand the importance of different variables.



1. **Which of the following is a disadvantage of non-parametric machine learning algorithms?**
    - a) Capable of fitting a large number of functional forms (flexibility)
    - b) Very fast to learn (speed)
    - c) More of a risk to overfit the training data (overfitting)
    - d) They do not require much training data

    Answer - c) More of a risk to overfit the training data (overfitting)
1. **Which of the following is a true statement for regression methods the in case of feature selection?**
    - a) Ridge regression uses subset selection of features
    - b) Lasso regression uses subset selection of features
    - c) Both use subset selection of features
    - d) None of above

    Answer - b) Lasso regression uses subset selection of features
1. **In Ridge regression, as the regularization parameter increases, do the regression coefficients decrease?**
    - a) True
    - b) False

    Answer - a) True
1. **Which regularization is used to reduce the over fit problem?**
    - a) L1
    - b) L2

- c) Both
- d) None of the above

Answer - a) L1

1. **Statement 1: The cost function is altered by adding a penalty equivalent to the square of the magnitude of the coefficients**
   Statement 2: Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent overfitting which may result from simple linear regression.
   - a) Statement 1 is true and statement 2 is false
   - b) Statement 1 is False and statement 2 is true
   - c) Both Statement (1 & 2) is true
   - d) Both Statement (1 & 2) is wrong

   Answer - c) Both Statement (1 & 2) is true
1. **Which of the following of the coefficients is added as the penalty term to the loss function in Lasso regression?**
   - a) Squared magnitude
   - b) Absolute value of magnitude
   - c) Number of non-zero entries
   - d) None of the above

   Answer - b) Absolute value of magnitude
1. **What type of penalty is used on regression weights in Ridge regression?**
   - a) L0
   - b) L1
   - c) L2
   - d) None of the above

   Answer - c) L2
1. **Lasso Regression uses which norm?**
   - a) L1.
   - b) L2.
   - c) L1 & L2 both.
   - d) None of the above.

   Answer - a) L1
1. **Ridge Regression uses which norm?**
   - a) L1.
   - b) L2.
   - c) L1 & L2 both.
   - d) None of the above.

   Answer - b) L2

1. **In Ridge regression, A hyper parameter is used called "_____" that controls the weighting of the penalty to the loss function.**

    - a) Alpha.
    - b) Gamma.
    - c) Lambda.
    - d) None of above.

    Answer - c) Lambda

1. **Ridge regression can reduce the slope close to zero (but not exactly zero) but Lasso regression can reduce the slope to be exactly equal to zero.**

    - a) Both statements are True about Ridge and Lasso.
    - b) Both statements are False about Ridge and Lasso.
    - c) True statement about Ridge but not about Lasso.
    - d) True statement about Lasso but not about Ridge.

    Answer - a) Both statements are True about Ridge and Lasso

1. **Ridge regression takes _____ value of variables.**

    - a) Squared value of variables.
    - b) Absolute value of variables.
    - c) Cube value of variables.
    - d) Root value of variables.

    Answer - a) Squared value of variables

1. **Ridge regression takes _____ value of variables.**

    - a) Squared value of variables.
    - b) Absolute value of variables.
    - c) Cube value of variables.
    - d) Root value of variables.

    Answer - b) Absolute value of variables

1. **The following statement is**

    I. Lasso Regression helps to reduce overfitting and it is particularly useful for feature selection.
    II. Lasso regression can be useful if we have several independent variables that are useless.

    - a) Statement (I) is true and statement (II) is false.
    - b) Statement (I) is false and statement (II) is true.
    - c) Both Statement (I) & (II) are wrong.
    - D) Both Statement (I) & (II) are true.

    Answer - D) Both Statement (I) & (II) are true

- **There should be no Outliers in the data points.**

Solution Approaches:

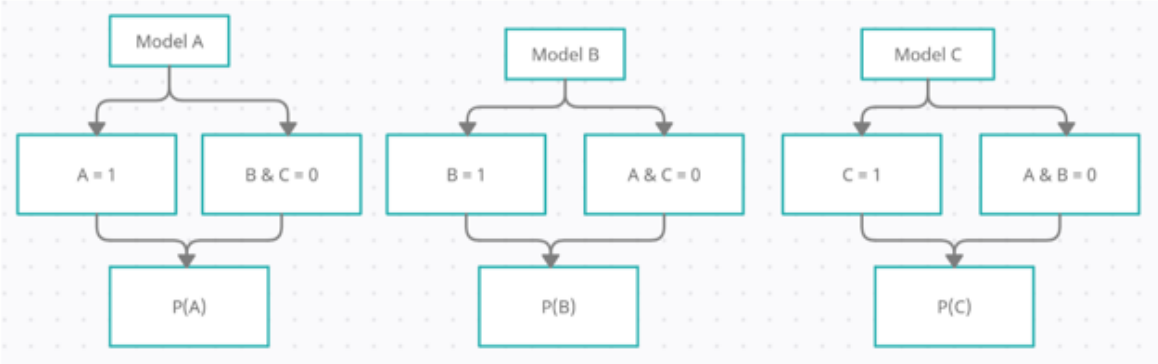
- **K models for K classes.**

This is the simplest approach where k models will be built for k classes as a set of independent binomial logistic regression.

For Example, there are three classes in nominal dependent variable i.e., A, B and C. Firstly, Build three models separately i.e. Class A vs Class B & C, Class B vs Class A & C and Class C vs Class A & B.

During First model, (Class A vs Class B & C): Class A will be 1 and Class B&C will be 0. In second model (Class B vs Class A & C): Class B will be 1 and Class A&C will be 0 and in third model (Class C vs Class A & B): Class C will be 1 and Class A&B will be 0.

Next develop the equation to calculate three Probabilities i.e. P(A), P(B) and P(C), very similar to the logistic regression equation.



Predicting the class of any record/observations, based on the independent input variables, will be the class that has highest probability. For a record, if P(A) > P(B) and P(A) > P(C), then the dependent target class = Class A.

- **Simultaneous Models.**

For K classes/possible outcomes, we will develop K-1 models as a set of independent binary regressions, in which one outcome/class is chosen as

“Reference/Pivot” class and all the other K-1 outcomes/classes are separately regressed against the pivot outcome.

When K = two, one model will be developed and multinomial logistic regression is equal to logistic regression.

For two classes i.e. Class A and Class B, one logistic regression model will be developed and the equation for probability is as follows:

$$Log\left(\frac{p}{1-p}\right) = a + b1x1 + b2x2 + \ldots + bnxn$$

If the value of p >= 0.5, then the record is classified as class A, else class B will be the possible target outcome.

For Multi-class dependent variables i.e. for K classes, K-1 Logistic Regression models will be developed. Let’s say there are three classes in dependent variable/Possible outcomes i.e. Class A, B and C.

Since there are three classes, two logistic regression models will be developed and let’s consider Class C has the reference or pivot class.

First Model will be developed for Class A and the reference class is C, the probability equation is as follows:

$$\left(\frac{p(A)}{p(C)}\right) = a1 + b1x1 + \ldots + bnxn$$

$$\frac{p(A)}{p(C)} = exp(a1 + b1x1 + \ldots + bnxn)$$

$$p(A) = p(C) * exp(a1 + b1x1 + \ldots + bnxn)$$

Develop second logistic regression model for class B with class C as reference class, then the probability equation is as follows:

$$\left(\frac{p(B)}{p(C)}\right) = a2 + b1x1 + \ldots + bnxn$$

$$\frac{p(B)}{p(C)} = exp(a2 + b1x1 + \ldots + bnxn)$$

$$p(B) = p(C) * exp(a2 + b1x1 + \ldots + bnxn)$$

Since P(A) +P(B) +P(C) = 1, then

$$p(C) * \exp exp\ (a1 + b1x1 + \ldots + bnxn) + p(C) * \exp exp\ (a2 + b1x1 + \ldots + bnxn) + p(C) = 1$$

$$p(C) = \frac{1}{1+exp(a1+b1x1+\ldots+bnxn)+\exp exp\ (a2+b1x1+\ldots+bnxn)}$$

Once probability of class C is calculated, probabilities of class A and class B can be calculated using the earlier equations.

Same logic can be applied to k classes where k-1 logistic regression models should be developed.

There are other approaches for solving the multinomial logistic regression problems.

Advantages:

– Helps to understand the relationships among the variables present in the dataset.

– Simultaneous Models result in smaller standard errors for the parameter estimates than when fitting the logistic regression models separately.

– The choice of reference class has no effect on the parameter estimates for other categories.