1. What is the difference between mean, median and mode? Application of each?

**Ans-** Averages are statistical constants which enable us to understand in a single effort the significance of the whole. They give us an idea about the concentration of the values in the central part of the distribution.

**Mean:** if we have a set of observations then our interest will be, what is the mean number of whole observations? The answer is that the sum of all observation divided by their total. It tells the average number.

Ex: 12, 15, 18,9,10,17,20,15 Mean= (12+15+18+9+10+17+20+15)/7

=14.5

**Median:** median divides the observations into two equal parts. In the scene of distribution it divides the distribution into two equal parts. It is the value such that the number of observations above it is equal to the number of observations below it. Thus the median is a position average.

Ex: 12, 15, 18, 9, 10, 17, 20, 15 median is the middle term i.e. (9+10)/2 =9.5

**Mode:** mode is the value which occurs most frequently in a set of observations. For example the average height of an Indian is 5-6. The average size of the shoes sold in a shop is 7.

An average student in a hostel spends Rs 1500 per month.

So the mode is the value which dominates the all values in series.

Ex: 12, 15, 18, 9, 10, 17, 20, 15 mode is the most frequently occurring value. That is in our example 15 repeated two times. Therefore mode is 15.

1. We already have variance and why the need of standard deviation came. Where we use these exactly. Cant one is enough?

**Ans-** the Standard Deviation is a measure of how spreads out numbers are. The average of the **squared** differences from the Mean is referred as variance.

Using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small. It may also be pointed out that standard deviation gives greater weight to extreme values and as such has not found favours with economists or businessmen who are not interested in the results of the model class. We may regard standard deviation as the best and the most powerful measure of dispersion.

1. If I give you a data what are the basic stats you provide me to understand the data behaviour?

**Ans-:** that is depending on the provided data, if it is grouped or ungrouped data we check its mean, median, mode, standard deviation, dispersion for knowing its central tendency. And we will check skewness and kurtosis for how the data distributed. If the given data is time series we will check for diagnosis.

And in case of regression data initially we will check whether it is following normality or not by plotting the data. If it is not normal we transform the data by suitable transformations. Then we see for multicollinearity, heteroschedasticity and autocorrelation etc.

1. Which method we use to fit the regression model? How many methods are there to fit regression model?

**Ans: -** To estimate the parameters of the linear regression model the most commonly used method of estimation is least squares method and for Non Linear Regression “maximum likelihood method”.

1. How do u find parameters of GLM (OLSE or MLE)

MLE

1. Why do we minimize squares of deviations (OLSE) why cannot we use absolute differences

**Ans:** - Because it’s hard to deal with absolute diff when you are differentiating and integrating.

1. What is MLE (maximum likelihood estimator)

**Ans:-** Used in estimating statistical parameters, it assumes a no distribution of the parameter and maximizes its joint probability distribution; estimate is obtained at the point where probability distribution of parameter is maximum. It assumes the distribution of the variable under consideration and finds out the parameters of that distribution by maximizing the likelihood function (JDF).

1. What is p-value

**Ans:-** It is the minimum level of significance which you reject the null hypothesis. OR More simply, If P value is less than 0.05 then reject null hypothesis and vice versa

1. When do u reject null hypothesis based on p-value

**Ans-:** When p value is less than level of significance (one tailed test)

When p value is less than ½\*level of significance (two tailed test)

1. What is the diff. between cov and corr

**Ans: -** Both are one and the same, corr is scaled version of cov (correlation is unit free, can be directly used in comparisons like corr(X1,Y1) > corr(X2,Y2), but it can’t be done directly with covariance)

1. Difference between correlation and regression

|  |  |
| --- | --- |
| **Correlation** | Linear **Regression** |
| Correlation examines the relationship between two variables using a [standardized unit](http://www.documentingexcellence.com/stat_tool/z-score.htm#standardized).  However, most applications use raw units as an input. | Regression examines the relationship between one dependent variables and one or more independent variables.  Calculations may us either raw unit values, or [standardized unit](http://www.documentingexcellence.com/stat_tool/z-score.htm#standardized)s as input. |
| The calculation is symmetrical, meaning that the order of comparison does NOT change the result. | The calculation is NOT symmetrical.  So one variable is assigned the dependent role (the values being predicted) and one or more the independent role (the values hypothesize to impact the dependent variable). |
| Correlation coefficients indicate the strength of a relationship. | Regression shows the effect of one unit change in an independent variable on the dependent variable. |
| Correlation removes the effect of different measurement scales.  Therefore, comparison between different models is possible since the rho coefficient is in standardized units. | Linear regression using raw unit measurement scales can be used to predict outcomes.  For example, if a model shows that spending more money on advertising will increases sales, then one can say that for every added $ in advertising our sales will increase by β. |

1. When you go for generalized linear models (GLM)?

**Ans:-** When Linear Regression model’s assumptions failed…….e.g.: errors are not normal;

OR When you have discrete independent variable

1. What are the assumptions of Linear Regression?

**Ans-** a) there should be linear relationship between dependent and independent variable.

b) Homoscaticity-Equal Variances between the groups.

c) No autocorrelation- observations should not depends on time.

d) Errors should be distributed i.i.d N(0,б\*\*2) i.e. identically, independently Normal distribution with mean zero and constant variance sigma square

1. What are the different kinds of outlier methods?

Ans: Observations which are different from all other can make a large difference in the results of regression analysis.

Outliers play important role in regression analysis. It is divided into two types of outliers one is that outlier in the response variable represent model failure. Such observations called outliers. And outliers with respect to predictions called the leverage points. They can affect the regression model too. In regression we can make two types of leverage points, good and bad leverage points.

A good leverage point is a point that is unusually large or small among the prediction values but it is not an outlier. That is we can remove that point from the predictors. A good leverage point has limited effect. It improves the accuracy in the model.

A bad leverage point is a point situated far from the regression line. That is it falls far from the all observations where all observations falls near to the regression line. Bad leverage point has an effect of the slope of the regression line. Bad leverage points reduce the accuracy of the model.

1. Describe Variable significance method in simple linear regression model.

Ans: we use the hypothesis test to determine whether there is a significance linear relationship between an independent variable “X” and dependent variable “Y” . The test focuses on the slope of the regression line Y=B0+B1X.

**Test requirements**:

The dependent variable Y has a linear relationship to the independent variable X.

For each value of X the probability distribution of Y has the same standard deviation “sigma”.

For any given value of X , the Y values are independent and the Y values are roughly normal distribution.

The null hypothesis states that the slope equal to zero. And the alternative hypothesis states that the slope is not equal to zero.

If the significance linear relationship between the independent variable X and dependent variable Y , Then the slope will not equal to zero.

H0:B1=0

H1:B1!=0.

1. What is ANOVA? Two applications.

**Ans:** In statistics analysis of variance is a collection of statistical models. In statistics ANOVA provides a statistical test whether the means of several groups are equal or not?

We can compare the means of several groups by conducting the generalized t- test. ANOVAs are helpful because they possess an advantage over a two sample t-test. There is a chance to committing type-1 error by doing multiple two sample t-test. For this reason, ANOVAs are useful in comparing two, three or more means.

1. Linearity assumptions:
   1. Linearity of the regression function ( scatter plot : residual vs. X)
   2. Independence of errors ( sequence plot)
   3. Constant variance of errors (residual vs. X)
   4. Normality of errors (Histogram of residuals , qq plot, Normal probability plot)
   5. These b, c, d together can say : errors follows normal distribution with mean 0 and same variance
2. When we transform the variables
3. When the basic assumptions(above) are not satisfied then go for transformations
4. What is R^2
5. Proportion of total variation in the response variable which has been explained by the explanatory variables.
6. Gives some information about the goodness of fit about the model
7. What is adjusted R^2
8. Adjusted R^2 is the modification of R^2 that adjust s the no of explanatory variables in the model. Unlike R^2.
9. Adjusted R^2 increases only if the new variables improves the model, more than would be expected by chance. But not R^2.
10. Adjusted R^2 can be negative, will always less than or equal to the R^2
11. What are the measures of checking model goodness of fit
12. R^2 and adjusted R^2
13. Each variable significance by using t-test with the p-value cut off 0.05
14. Model selection measures:
15. Mallows Cp criterion ( it should be small and approximately equal to no of explanatory variables) that is E (Cp) = P
16. AIC and SBC: don’t worry about formulas here: prefer small values of AIC and SBC. Since these are the functions of errors. When error is less then AIC and SBC are also less.
17. Outlier removal analysis: influential observations
18. One way can use sigma method: that is removing the observations which are above M+3\*SD or M-3\*SD. Again selecting this cut off depends on the data
19. Leverage > 2p/n, DFFITS and COOK’s distance
20. What is a Multi co linearity:
21. Variables are highly correlated
22. Poor value of R^2
23. Estimates will be unstable
24. Can be detected by using correlation matrix
25. Vif and tolerance : in general industry cut off is 10 for vif
26. Have to know the functionality of the vif
27. Multicollinearity issue can be resolve by- Drop the correlated variable, variable respecification, Ridge Regression method
28. What are the methods of selecting the best variables into the model:
    1. Forward
    2. Backward
    3. Stepwise ( read all these how it works)
    4. Stepwise is the most commonly used one in the industry
29. What is the difference between binomial and negative binomial distribution?
30. What is null hypothesis and alternative hypothesis
31. What false positive and false negative rate? ( need explain in general)
32. When we use logistic regression?( when Y is binary)
33. When we go for GLM? ( if errors do not follow normal distribution)
34. What is the distribution of dependant variable in the logistic regression? (Bernoulli )
35. What tests that check the significance of the variables in the logistic regression?( Wald test)
36. What is the concordance and discordance how this is useful in logistic regression?

* **Percent Concordant:** Percentage of pairs where the observation with the desired outcome (event) has a higher predicted probability than the observation without the outcome (non-event).
* **Percent Discordant:** Percentage of pairs where the observation with the desired outcome (event) has a lower predicted probability than the observation without the outcome (non-event).
* **Percent Tied:** Percentage of pairs where the observation with the desired outcome (event) has same predicted probability than the observation without the outcome (non-event).

In general, higher percentages of concordant pairs and lower percentages of discordant and tied pairs indicate a more desirable model.

1. What is the specificity and sensitivity?
2. What is the odds ratio?

Refer- <http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm>

1. Which method is used in logistic regression to fit the model?
2. What is the problem if we use binary variable in linear regression?
3. Why don’t we add error term in the logistic model?
4. What is logit?
5. What are the measures of to see logistic model goodness of fit?
   1. Maximum likelihood analysis
   2. Significance of each variable by using wald test
   3. Concordance ( should be high) and discordant (should be low)
6. Is multi co linearity creates any problem in logistic. If so then how it will be removed?
   1. Yes. In every model building it creates problem.
   2. Same by using vif we can remove.( note when u r applying vif we will not use response variable)
   3. Or by looking correlations also we can remove.
7. What is the bays theorem?
8. What is clustering? Where it will be useful?
9. What are the clustering methods? ( proc cluster And proc fastclus)
10. Explain end to end model building exercise.
11. Why train and test data is needed and what method widely uses to spate data for train and test?
12. What are the potential modelling approaches to predict dichotomous output?
13. What is confusion Matrix?
14. When CHAID is more effective than logistic regression?
15. What is the role of information gain in decision tree?
16. What type of support vector machine methods are there and which one best