1. What is applied econometrics?

Applied econometrics, on the other hand, concerns using economic metrics and measurements in a functional way. This might include studying economic benchmarks over a period of time to uncover trends or analyzing a set of specific metric points across several markets to determine probable outcomes in a given set of circumstances.

2. How can Econometrics be used as a tool for forecasting and prediction?

In the simplest terms, econometricians measure past relationships among such variables as consumer spending, household income, tax rates, interest rates, employment, and the like, and then try to forecast how changes in some variables will affect the future course of others.

3. What are the main steps in methodology of econometrics?

(i) Statement of theory or hypothesis. (ii) Specification of the mathematical model of the theory (iii) Specification of the statistical, or econometric, model (iv) Obtaining the data (v) Estimation of the parameters of the econometric model (vi) Hypothesis testing (vii) Forecasting or prediction (viii) Using the model for control or policy purposes.

4. Why do we need regression analysis? Why not simply use the mean value of the regress and as its best value?

In actual data we never get the mean of the regressand. We rather get samples which may be close or far away from the mean so it may be uncertain as to the sample value.

5.Give the two-variable line equation model. What do you understand by Stochastic Nature?

The solution of linear equations in two variables, ax + by = c, is a particular point in the graph, such that when x-coordinate is multiplied by a and y-coordinate is multiplied by b, then the sum of these two values will be equal to c.

6. What are the properties of GLS estimators and how to test the hypothesis for it?

The GLS is an unbiased estimator:

$E(\hat{}\beta GLS) = \beta + (X*TX*) - 1X*TE(\varepsilon*) = \beta E(\beta\hat{}GLS) = \beta + (X*TX*) - 1X*TE(\varepsilon*) = \beta$

The GLS variance-covariance matrix is:

$Var(\hat{}\beta GLS) = \sigma 2(X*TX*) - 1 = \sigma 2(XT\Psi\Psi TX) - 1 = \sigma 2(XT\Omega - 1X) - 1 Var(\beta\hat{}GLS) = \sigma 2(X*TX*) - 1 = \sigma 2(XT\Psi\Psi TX) - 1 = \sigma 2(XT\Omega - 1X) - 1$

If the errors are normally distributed, then: ˆ

$\beta GLS|X \sim N(\beta, \sigma 2(XT\Omega - 1X) - 1)$ $\beta\hat{}GLS|X \sim N(\beta, \sigma 2(XT\Omega - 1X) - 1)$

$\hat{}\sigma 2\sigma\hat{}2$ is unbiased and consistent.

7. What are the properties of least square estimators?

The basic distributional assumptions of the linear model are

(a) The errors are unbiased: $E[\varepsilon] = 0$.

(b) The errors are uncorrelated with common variance: $cov(\varepsilon) = \sigma 2I$.

These assumptions imply that $E[Y] = X\beta$ and $cov(Y) = \sigma 2I$.

8. What are the assumptions of the linear regression model?

There are four assumptions associated with a linear regression model:

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

9. What are the criteria on which the estimated model is to be evaluated?

Three statistics are used in Ordinary Least Squares (OLS) regression to evaluate model fit: R-squared, the overall F-test, and the Root Mean Square Error (RMSE). All three are based on two sums of squares: Sum of Squares Total (SST) and Sum of Squares Error (SSE).

10. Explain the regression approach of analyzing qualitative choice situations.

Regression analysis is a reliable method of identifying which variables have impact on a topic of interest. The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other.

In order to understand regression analysis fully, it's essential to comprehend the following terms:

Dependent Variable: This is the main factor that you're trying to understand or predict.

Independent Variables: These are the factors that you hypothesize have an impact on your dependent variable.

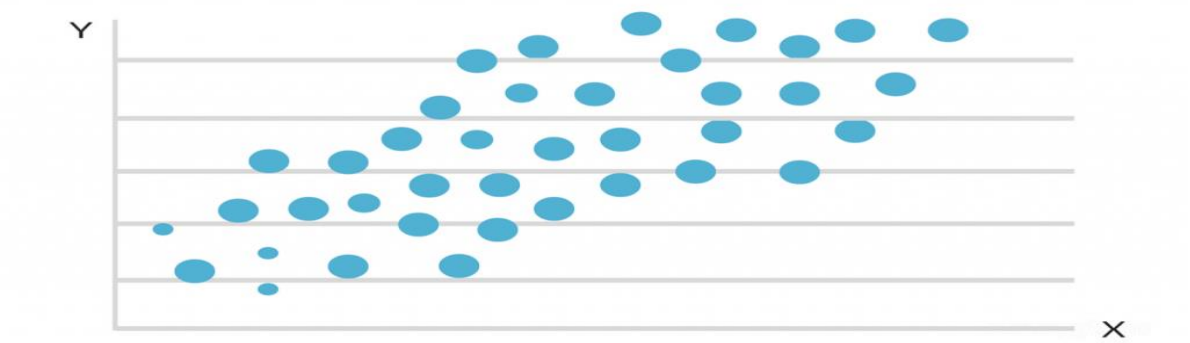
In our application training example above, attendees' satisfaction with the event is our dependent variable. The topics covered, length of sessions, food provided, and the cost of a ticket are our independent variables.

In order to conduct a regression analysis, you'll need to define a dependent variable that you hypothesize is being influenced by one or several independent variables.

You'll then need to establish a comprehensive dataset to work with. Administering surveys to your audiences of interest is a terrific way to establish this dataset. Your survey should include questions addressing all of the independent variables that you are interested in.

Let's continue using our application training example. In this case, we'd want to measure the historical levels of satisfaction with the events from the past three years or so (or however long you deem statistically significant), as well as any information possible in regards to the independent variables. Perhaps we're particularly curious about how the price of a ticket to the event has impacted levels of satisfaction.

To begin investigating whether or not there is a relationship between these two variables, we would begin by plotting these data points on a chart, which would look like the following theoretical example.



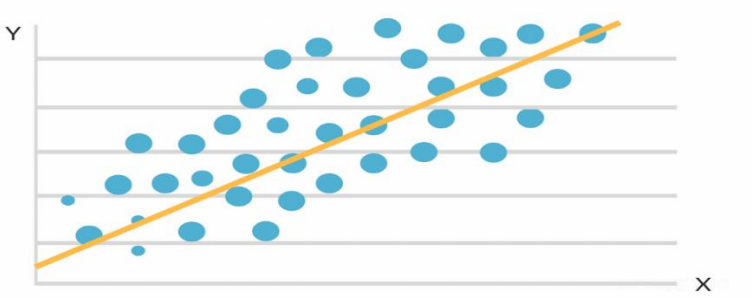
(Plotting your data is the first step in figuring out if there is a relationship between your independent and dependent variables)

Our dependent variable (in this case, the level of event satisfaction) should be plotted on the y-axis, while our independent variable (the price of the event ticket) should be plotted on the x-axis. Once your data is plotted, you may begin to see correlations. If the theoretical chart above did indeed represent the impact of ticket prices on event satisfaction, then we'd be able to confidently say that the higher the ticket price, the higher the levels of event satisfaction.

But how can we tell the degree to which ticket price affects event satisfaction?

To begin answering this question, draw a line through the middle of all of the data points on the chart. This line is referred to as your regression line, and it can be precisely calculated using a standard statistics program like Excel.

We'll use a theoretical chart once more to depict what a regression line should look like.



The regression line represents the relationship between your independent variable and your dependent variable.
Excel will even provide a formula for the slope of the line, which adds further context to the relationship between your independent and dependent variables.
The formula for a regression line might look something like Y = 100 + 7X + error term.
This tells you that if there is no "X", then Y = 100. If X is our increase in ticket price, this informs us that if there is no increase in ticket price, event satisfaction will still increase by 100 points.
You'll notice that the slope formula calculated by Excel includes an error term. Regression lines always consider an error term because in reality, independent variables are never precisely perfect predictors of dependent variables. This makes sense while looking at the impact of ticket prices on event satisfaction — there are clearly other variables that are contributing to event satisfaction outside of price.
Your regression line is simply an estimate based on the data available to you. So, the larger your error term, the less definitively certain your regression line is.

Regression analysis is helpful statistical method that can be leveraged across an organization to determine the degree to which particular independent variables are influencing dependent variables.
The possible scenarios for conducting regression analysis to yield valuable, actionable business insights are endless.
The next time someone in your business is proposing a hypothesis that states that one factor, whether you can control that factor or not, is impacting a portion of the business, suggest performing a regression analysis to determine just how confident you should be in that hypothesis! This will allow you to make more informed business decisions, allocate resources more efficiently, and ultimately boost your bottom line.

11. Estimate the regression line with its slope & intercept.

Consumption (Y) (Rs)	70	65	90	95	110	115	120	140	145	150
Income (X) (Rs)	80	100	120	140	160	180	200	220	240	260

12. What is meant by the dummy variable trap?
The Dummy Variable Trap occurs when two or more dummy variables created by one-hot encoding are highly correlated (multi-collinear). This means that one variable can be predicted from the others, making it difficult to interpret predicted coefficient variables in regression models. In other words, the individual effect of the dummy variables on the prediction model cannot be interpreted well because of multicollinearity.
13. How many dummy variables?
The number of dummy variables required to represent a particular categorical variable depends on the number of values that the categorical variable can assume. To represent a categorical variable that can assume k different values, a researcher would need to define k - 1 dummy variables.
For example, suppose we are interested in political affiliation, a categorical variable that might assume three values - Republican, Democrat, or Independent. We could represent political affiliation with two dummy variables:
X_1 = 1, if Republican; X_1 = 0, otherwise.
X_2 = 1, if Democrat; X_2 = 0, otherwise.

In this example, notice that we don't have to create a dummy variable to represent the "Independent" category of political affiliation. If X_1 equals zero and X_2 equals zero, we know the voter is neither Republican nor Democrat. Therefore, voter must be Independent.

14. **How to deal with multicollinearity?**

If you only want to predict the value of a dependent variable, you may not have to worry about multicollinearity. Multiple regression can produce a regression equation that will work for you, even when independent variables are highly correlated. The problem arises when you want to assess the relative importance of an independent variable with a high R^2_k (or, equivalently, a high VIF_k). In this situation, try the following:

- Redesign the study to avoid multicollinearity. If you are working on a true experiment, the experimenter controls treatment levels. Choose treatment levels to minimize or eliminate correlations between independent variables.
- Increase sample size. Other things being equal, a bigger sample means reduced sampling error. The increased precision may overcome potential problems from multicollinearity.
- Remove one or more of the highly-correlated independent variables. Then, define a new regression equation, based on the remaining variables. Because the removed variables were redundant, the new equation should be nearly as predictive as the old equation; and coefficients should be easier to interpret because multicollinearity is reduced.
- Define a new variable equal to a linear combination of the highly-correlated variables. Then, define a new regression equation, using the new variable in place of the old highly-correlated variables.

Note: Multicollinearity only affects variables that are highly correlated. If the variable you are interested in has a small R^2_j, statistical analysis of its regression coefficient will be reliable and informative. That analysis will be valid, even when other variables exhibit high multicollinearity.

15. **What is perfect multicollinearity, state the effects of it.**

Perfect (or Exact) Multicollinearity If two or more independent variables have an exact linear relationship between them then we have perfect multicollinearity. Examples: including the same information twice (weight in pounds and weight in kilograms), not using dummy variables correctly (falling into the dummy variable trap), etc.

16. **What are the consequences of multicollinearity?**

The Consequences of Multicollinearity

1. Imperfect multicollinearity does not violate Assumption 6.

Therefore, the Gauss Markov Theorem tells us that the OLS estimators are BLUE. So then why do we care about multicollinearity?

2. The variances and the standard errors of the regression coefficient estimates will increase. This means lower t-statistics.

3. The overall fit of the regression equation will be largely unaffected by multicollinearity. This also means that forecasting and prediction will be largely unaffected.

4. Regression coefficients will be sensitive to specifications. Regression coefficients can change substantially when variables are added or dropped.

17. Explain the Breusch-Pagan test of heteroscedasticity.

It is used to test for heteroskedasticity in a linear regression model and assumes that the error terms are normally distributed. It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. It is a $\chi 2$ test.

18. Give the figures for different forms of autocorrelation.

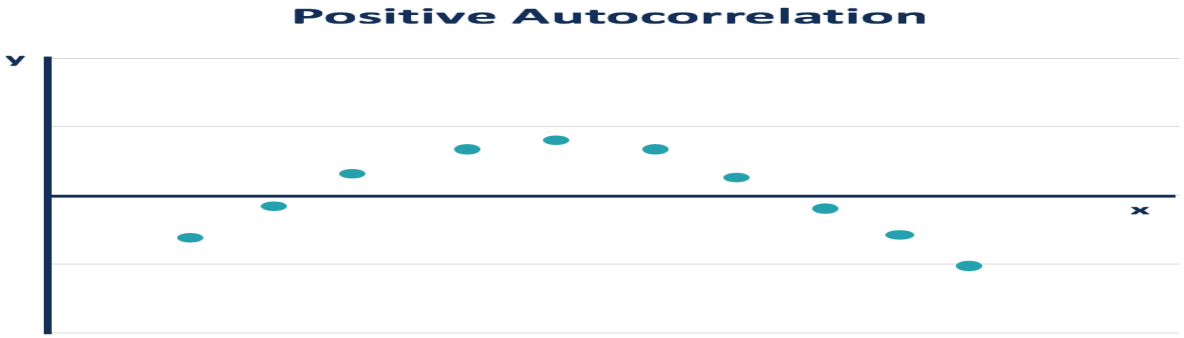
In many cases, the value of a variable at a point in time is related to the value of it at a previous point in time. Autocorrelation analysis measures the relationship of the observations between the different points in time, and thus

seeks a pattern or trend over the time series. For example, the temperatures on different days in a month are autocorrelated.

Similar to correlation, autocorrelation can be either positive or negative. It ranges from -1 (perfectly negative autocorrelation) to 1 (perfectly positive autocorrelation). Positive autocorrelation means that the increase observed in a time interval leads to a proportionate increase in the lagged time interval.

The example of temperature discussed above demonstrates a positive autocorrelation. The temperature the next day tends to rise when it's been increasing and tends to drop when it's been decreasing during the previous days.

The observations with positive autocorrelation can be plotted into a smooth curve. By adding a regression line, it can be observed that a positive error is followed by another positive one, and a negative error is followed by another negative one.



Conversely, negative autocorrelation represents that the increase observed in a time interval leads to a proportionate decrease in the lagged time interval. By plotting the observations with a regression line, it shows that a positive error will be followed by a negative one and vice versa.

Autocorrelation can be applied to different numbers of time gaps, which is known as lag. A lag 1 autocorrelation measures the correlation between the observations that are a one-time gap apart. For example, to learn the correlation between the temperatures of one day and the corresponding day in the next month, a lag 30 autocorrelation should be used (assuming 30 days in that month).

19. **What are the implications of autocorrelation?**

When autocorrelation is detected in the residuals from a model, it suggests that the model is mis specified (i.e., in some sense wrong). A cause is that some key variable or variables are missing from the model. Where the data has been collected across space or time, and the model does not explicitly account for this, autocorrelation is likely. For example, if a weather model is wrong in one suburb, it will likely be wrong in the same way in a neighboring suburb. The fix is to either include the missing variables, or explicitly model the autocorrelation (e.g., using an ARIMA model).

19.Explain the Durbin-Watson Test. What are its assumptions?

This test was developed by Statisticians Durbin and Watson.

It is most frequently used test for the detection of autocorrelation.

It is also called as Durbin– Watson d test.

It is used to test the null hypothesis that there is no autocorrelation.

The value of d statistic lies between 0 and 4.
A value near 0 shows the presence of positive autocorrelation, value near 4 shows presence of negative autocorrelation whereas value near two shows absence of autocorrelation.
However, it is difficult to decide how much near to 0, 2 or 4. Therefore a criterion was suggested by Durbin and Watson. They construct a table for upper bound (dU) and lower bound (dL) for the d statistic. This table is for 6 to 200 observations and maximum 20 explanatory variables. The decision criteria are explained in figure on next slide.

The Durbin – Watson test is used under following assumptions only:
1. The Regression model includes intercept term.
2. The explanatory variables must be non-stochastic (nonrandom or fixed).
3. The error term must be normally distributed.
4. The regression term doesn't include any lagged (past) value of dependent variable.
5. There must be no missing observation.
6. It can be used only for first order autocorrelation.

20. Give important details on the instrumental variable's method.
In statistics, econometrics, epidemiology and related disciplines, the method of instrumental variables (IV) is used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment.

21. Explain the Hausman specification test.
The Durbin–Wu–Hausman test (also called Hausman specification test) is a statistical hypothesis test in econometrics named after James Durbin, De-Min Wu, and Jerry A. Hausman. The test evaluates the consistency of an estimator when compared to an alternative, less efficient estimator which is already known to be consistent. It helps one evaluate if a statistical model corresponds to the data.
The Hausman test can be used to differentiate between fixed effects model and random effects model in panel analysis. In this case, Random effects (RE) is preferred under the null hypothesis due to higher efficiency, while under the alternative Fixed effects (FE) is at least as consistent and thus preferred.
22. Find the value of b_1 and b_2 using deviation method:

X	50	45	40	35
Y	10	15	20	25

23. What is the difference between panel data, time-serial data, and cross-sectional data?
Time series data - It is a collection of observations(behavior) for a single subject(entity) at different time intervals (generally equally spaced) Example - Max Temperature, Humidity and Wind (all three behaviors) in New York City (single entity) collected on First day of every year (multiple intervals of time)

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2012	35	56%	3 mph
NYC	1/1/2013	47	65%	21 mph
NYC	1/1/2014	30	39%	16 mph
NYC	1/1/2015	55	45%	4 mph

Cross-Sectional data - It is a collection of observations(behavior) for multiple subjects(entities) at single point in time. Example - Max Temperature, Humidity and Wind (all three behaviors) in New York City, SFO, Boston, Chicago (multiple entities) on 1/1/2015(single instance)

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2015	55	45%	4 mph
SFO	1/1/2015	70	35%	21 mph
Boston	1/1/2015	34	39%	16 mph
Chicago	1/1/2015	29	15%	54 mph

Panel Data (Longitudinal Data) - It is usually called as Cross-sectional Time-series data as it a combination of above-mentioned types, i.e., collection of observations for multiple subjects at multiple instances. Example - Max Temperature, Humidity and Wind (all three behaviors) in New York City, SFO, Boston, Chicago (multiple entities) on First day of every year (multiple intervals of time)

City	Date	MaxTemperature	Humidity	Wind
NYC	1/1/2015	55	45%	4 mph
NYC	1/1/2014	30	39%	16 mph
NYC	1/1/2013	47	65%	21 mph
SFO	1/1/2015	70	35%	21 mph
SFO	1/1/2014	75	23%	2 mph
SFO	1/1/2013	71	39%	13 mph
Boston	1/1/2015	34	39%	16 mph
Boston	1/1/2014	26	17%	27 mph
Boston	1/1/2013	45	46%	18 mph

28. **When to use a one-way ANOVA**

Use a one-way ANOVA when you have collected data about one categorical independent variable and one quantitative dependent variable. The independent variable should have at least three levels (i.e. at least three different groups or categories). ANOVA tells you if the dependent variable changes according to the level of the independent variable. For example:

- Your independent variable is social media use, and you assign groups to low, medium, and high levels of social media use to find out if there is a difference in hours of sleep per night.
- Your independent variable is brand of soda, and you collect data on Coke, Pepsi, Sprite, and Fanta to find out if there is a difference in the price per 100ml.
- Your independent variable is type of fertilizer, and you treat crop fields with mixtures 1, 2 and 3 to find out if there is a difference in crop yield.

The null hypothesis (H_0) of ANOVA is that there is no difference among group means. The alternate hypothesis (H_a) is that at least one group differs significantly from the overall mean of the dependent variable. If you only want to compare two groups, use a t-test instead.

29. **How does an ANOVA test work?**

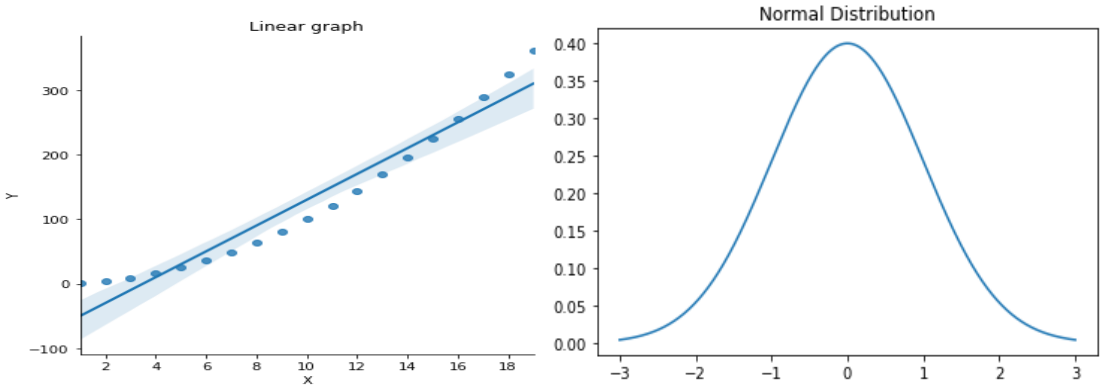
ANOVA determines whether the groups created by the levels of the independent variable are statistically different by calculating whether the means of the treatment levels are different from the overall mean of the dependent variable.

If any of the group means is significantly different from the overall mean, then the null hypothesis is rejected.

ANOVA uses the F-test for statistical significance. This allows for comparison of multiple means at once, because the error is calculated for the whole set of comparisons rather than for each individual two-way comparison (which would happen with a t-test).
The F-test compares the variance in each group mean from the overall group variance. If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-value, and therefore a higher likelihood that the difference observed is real and not due to chance.
The F-test compares the variance in each group mean from the overall group variance. If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-value, and therefore a higher likelihood that the difference observed is real and not due to chance.

30. What is a Generalized Linear Model?

Generalized Linear Model (GLM) is an advanced statistical modelling technique formulated by John Nelder and Robert Wedderburn in 1972. It is an umbrella term that encompasses many other models, which allows the response variable y to have an error distribution other than a normal distribution. The models include Linear Regression, Logistic Regression, and Poisson Regression.
In a Linear Regression Model, the response (aka dependent/target) variable 'y' is expressed as a linear function/linear combination of all the predictors 'X' (aka independent / regression / explanatory / observed variables).
The underlying relationship between the response and the predictors is linear (i.e. we can simply visualize the relationship in the form of a straight line). Also, the error distribution of the response variable should be normally distributed. Therefore, we are building a linear model.

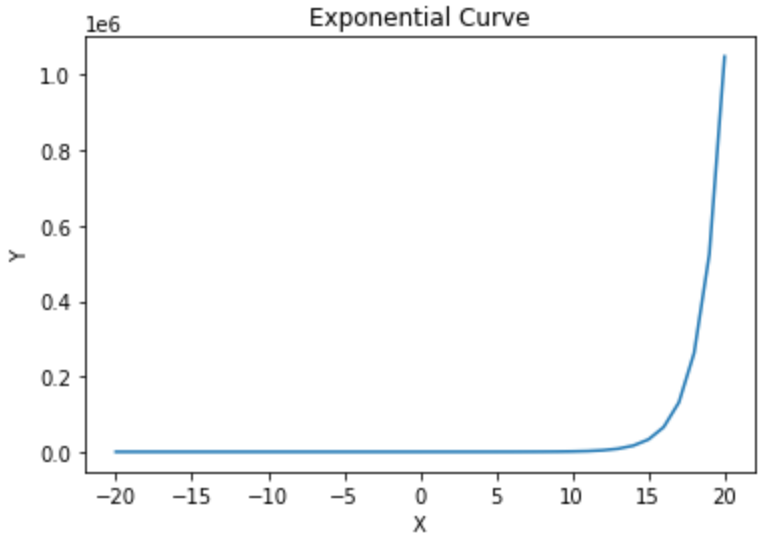


GLM models allow us to build a linear relationship between the response and predictors, even though their underlying relationship is not linear. This is made possible by using a link function, which links the response variable to a linear model. Unlike Linear Regression models, the error distribution of the response variable need not be normally distributed. The errors in the response variable are assumed to follow an exponential family of distribution (i.e. normal, binomial, Poisson, or gamma distributions). Since we are trying to generalize a linear regression model that can also be applied in these cases, the name Generalized Linear Models.

31. Why GLM?

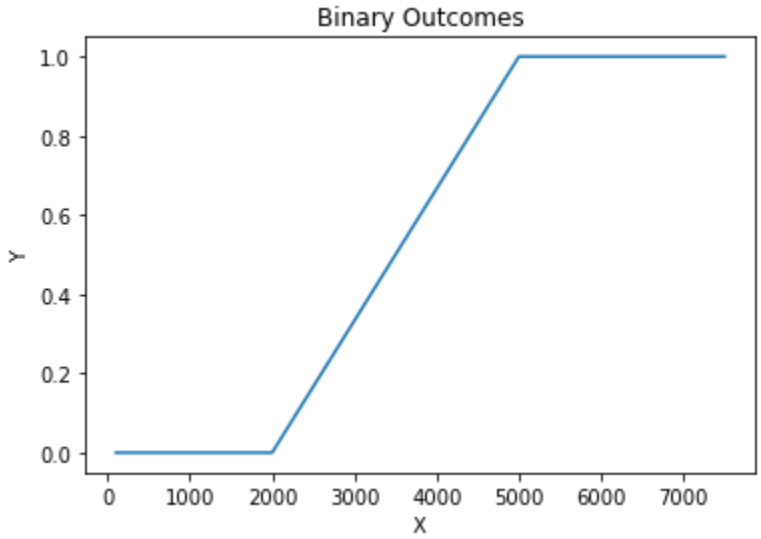
Linear Regression model is not suitable if,

- The relationship between X and y is not linear. There exists some non-linear relationship between them. For example, y increases exponentially as X increases.



- Variance of errors in y (commonly called as Homoscedasticity in Linear Regression), is not constant, and varies with X.
- Response variable is not continuous, but discrete/categorical. Linear Regression assumes normal distribution of the response variable, which can only be applied on a continuous data. If we try to build a linear regression model on a discrete/binary y variable, then the linear regression model predicts negative values for the corresponding response variable, which is inappropriate.

In the below graph, we can see the response is either 0 or 1. When X <5000, y is 0, and when X>=5000, y is 1



For Example, consider a linear model as follows:
A simple example of a mobile price in an e-commerce platform:
Price = 12500 + 1.5*Screen size – 3*Battery Backup (less than 4hrs)
Data available for,

- Price of the mobile
- Screen size (in inches)
- Is battery backup less than 4hrs – with values either as ‘yes’, or ‘no’.

In this example, if the screen size increases by 1 unit, then the price of the mobile increases by 1.5 times the default price, keeping the intercept (12500) and Battery Backup values constant. Likewise, if the Battery Backup of less than 4hrs is ‘yes, then the mobile price reduces by three times the default price. If the Battery

Backup of less than 4hrs is ‘no’, then the mobile price is unaffected, as the term (3*Battery Backup) becomes 0 in the linear model. The intercept 12500 indicates the default price for a standard value of screen size. This is a valid model.
However, if we get a model as below:
Price = 12500 +1.5*Screen size + 3*Battery Backup (less than 4hrs)
Here, if the battery backup less than 4 hrs is ‘yes, then the model is saying the price of the phone increases by three times. Clearly, from practical knowledge, we know this is incorrect. There will be less demand for such mobiles. These are going to be very old mobiles, which when compared to the current range of mobiles with the latest features, is going to be very less in price. This is because the relationship between the two variables is not linear, but we are trying to express it as a linear relationship. Hence, an invalid model is built.
Similarly, if we are trying to predict if a particular phone will be sold or not, using the same independent variables, but the target is we are trying to predict if the phone will sell or not, so it has only binary outcomes.
Using Linear Regression, we get a model like,
Sales = 12500 +1.5*Screen size – 3*Battery Backup (less than 4hrs)
This model doesn’t tell us if the mobile will be sold or not, because the output of a linear regression model is continuous value. It is possible to get negative values as well as the output. It does not translate to our actual objective of whether phones having some specifications based on the predictors, will sell or not (binary outcome).
Similarly, if we are also trying to see what is the number of sales of this mobile that will happen in the next month, a negative value means nothing. Here, the minimum value is 0 (no sale happened), or a positive value corresponding to the count of the sales. Having the count as a negative value is not meaningful to us.

32. Discuss assumptions of GLM

Similar to Linear Regression Model, there are some basic assumptions for Generalized Linear Models as well. Most of the assumptions are similar to Linear Regression models, while some of the assumptions of Linear Regression are modified.

- Data should be independent and random (Each Random variable has the same probability distribution).
- The response variable y does not need to be normally distributed, but the distribution is from an exponential family (e.g. binomial, Poisson, multinomial, normal)
- The original response variable need not have a linear relationship with the independent variables, but the transformed response variable (through the link function) is linearly dependent on the independent variables

Ex., Logistic Regression Equation, Log odds = $\beta 0+\beta 1X1+\beta 2X2$,
where $\beta 0,\beta 1,\beta 2$ are regression coefficient, and X1,X2 are the independent variables

- Feature engineering on the Independent variable can be applied i.e instead of taking the original raw independent variables, variable transformation can be done, and the transformed independent variables, such as taking a log transformation, squaring the variables, reciprocal of the variables, can also be used to build the GLM model.
- Homoscedasticity (i.e constant variance) need not be satisfied. Response variable Error variance can increase, or decrease with the independent variables.
- Errors are independent but need not be normally distributed.

33. Explain main components of GLM

There are 3 components in GLM.

- Systematic Component/Linear Predictor:

It is just the linear combination of the Predictors and the regression coefficients.

$\beta 0+\beta 1X1+\beta 2X2$

- Link Function:

Represented as η or $g(\mu)$, it specifies the link between a random and systematic component. It indicates how the expected/predicted value of the response relates to the linear combination of predictor variables.

- Random Component/Probability Distribution:

It refers to the probability distribution, from the family of distributions, of the response variable.
The family of distributions, called an exponential family, includes normal distribution, binomial distribution, or Poisson distribution.
Below summarizes the table of Probability Distribution, and their corresponding Link function

Probability Distribution	Link Function
Normal Distribution	Identity function
Binomial Distribution	Logit/Sigmoid function
Poisson Distribution	Log function (aka log-linear, log-link)

34. Explain Different Generalized Linear Models

Commonly used models in the GLM family include:
Linear Regression, for continuous outcomes with normal distribution:
Here we model the mean expected value of a continuous response variable as a function of the explanatory variables. Identity link function is used, which is the simplest link function.
If there is only 1 predictor, then the model is called Simple Linear Regression. If there are 2 or more explanatory variables, then the model is called Multiple Linear Regression.
Simple Linear Regression, $y= \beta 0+\beta 1X1$
Multiple Linear Regression, $y = \beta 0+\beta 1X1+\beta 2X2$
Response is continuous
Predictors can be continuous or categorical, and can also be transformed.
Errors are distributed normally and variance is constant.
Binary Logistic Regression, for dichotomous or binary outcomes with binomial distribution:
Here Log odds is expressed as a linear combination of the explanatory variables. Logit is the link function. The Logistic or Sigmoid function, returns probability as the output, which varies between 0 and 1.
Log odds= $\beta 0+\beta 1X1+\beta 2X2$
Response variable has only 2 outcomes
Predictors can be continuous or categorical, and can also be transformed.

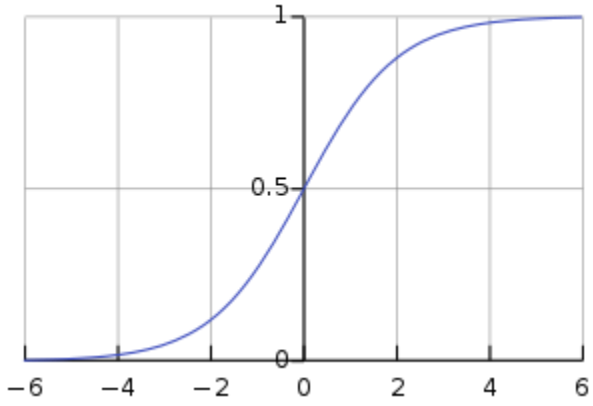
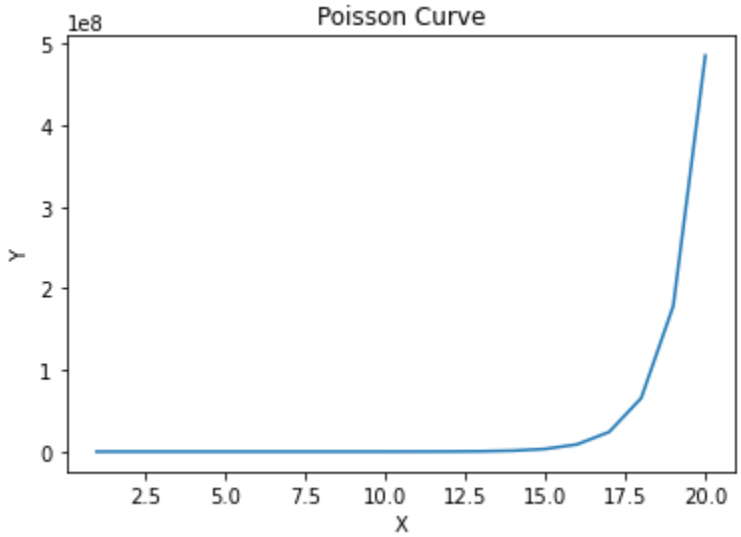


Image source: https://en.wikipedia.org/wiki/Sigmoid_function

Poisson Regression, for count based outcomes with Poisson distribution:
Here count values are expressed as a linear combination of the explanatory variables. Log link is the link function.
$\log(\lambda)=\beta 0+\beta 1\times 1+\beta 2\times 2,$
where λ is the average value of the count variable
Response variable is a count value per unit of time and space
Predictors can be continuous or categorical, and can also be transformed.



35. Difference Between Generalized Linear Model and General Linear Model

General Linear Models, also represented as GLM, is a special case of Generalized Linear Models (GLiM). General Linear Models refers to normal linear regression models with a continuous response variable. It includes many statistical models such as Single Linear Regression, Multiple Linear Regression, Anova, Ancova, Manova, Mancova, t-test and F-test. General Linear Models assumes the residuals/errors follow a normal distribution. Generalized Linear Model, on the other hand, allows residuals to have other distributions from the exponential family of distributions.

36. Can Generalized Linear Models have correlated data?

For Generalized Linear Models, data should not be correlated with each other. If the data is correlated, then the model performance will not be reliable. For this reason, GLMs are unsuitable on time series data, where usually data will have some auto-correlation in them. However, some variations of GLM have also been developed to consider the correlation in the data, such as the Generalized Estimating Equations (GEEs) model and Generalized Linear Mixed Models (GLMMs) model.