

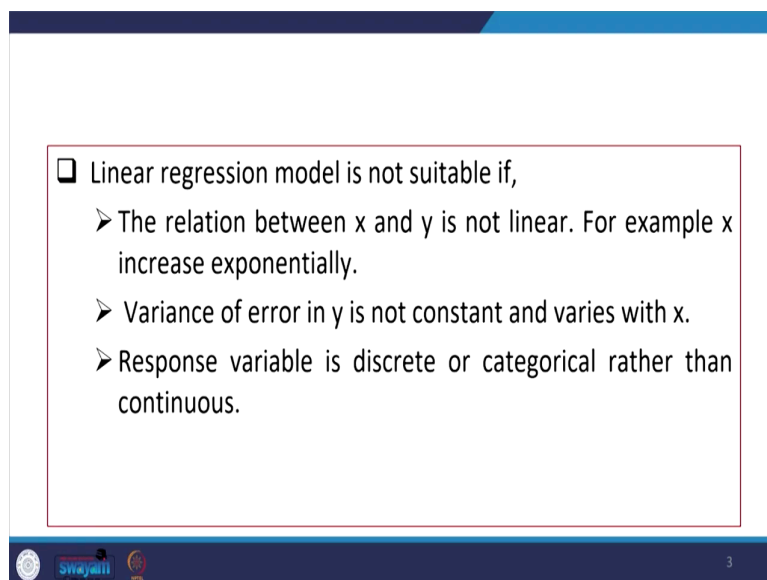
Exploring Survey Data on Health Care
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 29
Generalized Linear Model

Welcome friends once again to my NPTEL MOOC module on Exploring Healthcare Survey Data. We are in the last week of the module. We have been discussing about health care data and in this particular lecture we are trying to understand generalized linear model. This GLM is used when your linear model has certain limitations and all the assumptions are not fulfilled.


Then we go for some generalizations of that linear model. Now I am just going to discuss about who has developed it? & how it has emerged? So, GLM is an advanced statistical modelling technique that generalizes the linear model. It was developed by John Nelder and Robert Wedderburn in 1972. It is an umbrella term that encompasses many other models which allows the response variable i.e., Y , to have an error distribution other than a normal distribution

(Refer Slide Time: 01:34)



❑ Linear regression model is not suitable if,

- The relation between x and y is not linear. For example x increase exponentially.
- Variance of error in y is not constant and varies with x .
- Response variable is discrete or categorical rather than continuous.

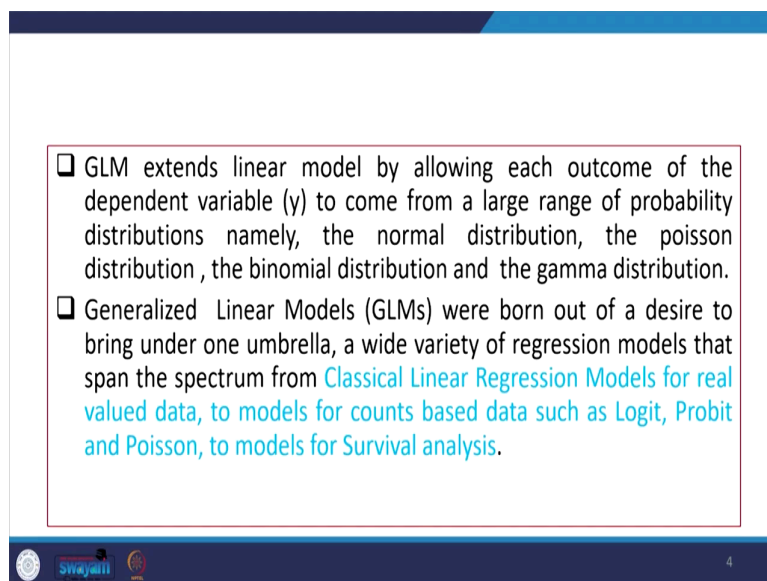
swayam  3

Linear regression model is not suitable if the relations between X and Y is not linear. For example, X increases exponentially, if there is exponential distribution like cumulative density function which is not linear. In fact, in that case linear regression is not suitable.

Similarly, another one is called variance of error in the Y i.e., Y is not constant and varies with X. So, the variance, if it is not constant throughout then in that case we already know that linear regression model is not fitted and we go for some form of transformation.

If Response variable is discrete or categorical rather than a continuous one then in that case linear regression is also not applicable. Therefore, we think of some possibility of some generalization and try to use generalized least square method or generalized least square model or generalized linear model.

(Refer Slide Time: 02:41)



- ❑ GLM extends linear model by allowing each outcome of the dependent variable (y) to come from a large range of probability distributions namely, the normal distribution, the poisson distribution, the binomial distribution and the gamma distribution.
- ❑ Generalized Linear Models (GLMs) were born out of a desire to bring under one umbrella, a wide variety of regression models that span the spectrum from [Classical Linear Regression Models for real valued data](#), to [models for counts based data such as Logit, Probit and Poisson](#), to [models for Survival analysis](#).

GLM extends linear model by allowing each outcome of the dependent variable to come from a large range of probability distribution. Namely, the normal distribution, the Poisson distribution, the binomial distribution and the gamma distribution. You might have heard about this distribution. It has certain specifications or certain features of Poisson. We know that it has certain count data of binomial, we know that it follows Bernoulli distribution with binary numbers. So, gamma functions are accordingly defined normal distribution where the data is of bell kind shape, symmetry is usually identified.

The generalized linear models were born out of a desire to bring under one umbrella, a wide variety of regression models that span the spectrum from classical linear regression models for real valued data to models that are counts based data such as logit, or probit or Poisson, to models for survival analysis etc. just to give another example.

Poisson is like if you have certain count data such as number of consultations of doctors, number of visits to the doctors etc., those are simply count and not continuous series. The number of consultations cannot be just 1.8, 1.9 or 1.2. There is no fraction possibility. It has to be a count or absolute number and therefore, they are called Poisson kind of data.

(Refer Slide Time: 04:24)

□ GLM is made up of a linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

and two functions

- A **link function** that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor $g(\mu_i) = \eta_i$
- a **variance function** that describes how the variance, $\text{var}(Y_i)$ depends on the mean $\text{var}(Y_i) = \phi V(\mu)$ where the dispersion parameter ϕ is a constant.

So, GLM is made of a linear predictor of course, that is mentioned with a function called beta naught (β^0) plus beta1_X1I plus up to beta_p Xi pi. And two functions are usually discussed in the case of GLM set up. One is called a link function, which describes how you get the mean or the expected mean with the population mean depends on the linear predictor i.e., the predictor value of the mean. And second one is the variance function, that describes how the variance that is the var of Yi depends on the mean value of it. And that can be of this value as per the equation whether that dispersion parameter theta is considered to be constant.

(Refer Slide Time: 05:20)

Component of GLM

- **Random component**
 - refers to the probability distribution of the dependent variable (Y).
 - E.g. normal distribution for Y in the linear regression or binomial distribution for Y in the binary logistic model.
- **Systematic component**

It specifies the explanatory ($x_1, x_2, x_3, \dots, x_k$) variables in the model as a combination of linear predictors.
- **Link function,**
 - Specifies the link between random and systematic components.
 - It says how the expected value of the response relates to the linear predictor of explanatory variables, as $\eta = g(E(Y_i)) = E(Y_i)$ for linear regression and $\eta = \text{logit}(\pi)$ for logistic regression

swayamii 6

So, some of the components those are usually discussed in GLM are called random component, systematic component and link function. Random component is something that refers to the probability distribution of the dependent variable. For example, normal distribution for Y in the linear regression or binomial distribution for Y in the binary logistic function. In case of systematic component, it specifies the explanatory variable such as X1 to Xk variables in the model as a combination of linear predictors.

Link function specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables. And these are given in this equation with the function expected value of Y, how the function further transformation of that function is also related to the expected value of Y for linear regression. And that can be also applied through the logit function of the pi that is discussed accordingly we shared it as logistic regression

(Refer Slide Time: 06:34)

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

So here is a summary of GLM following Agresti paper of chapter 4 published in 2013. It has given the GLM functions or models and how it is applied. As we know, the linear regression is applied when the data is continuous.

But an ANOVA or ANCOVA analysis of covariance or variance are also discussed in the context of categorical and mixed data. The link function gives the identity of it, the identity of the variable that has discussed the randomness, is in fact defined to be normal. The distribution is defined to be normal. So, in case of the specific GLM model like logistic loglinear, Poisson and multinomial response.

The random number table or randomness of this distribution is no more normal, they are either binomial Poisson or multinomial and according the link function it is defined as logit log function or generalized logit function. And therefore, it has mixed responses.

(Refer Slide Time: 07:50)

Linear Model	Generalized linear model
<ul style="list-style-type: none"><input type="checkbox"/> In linear regression model the dependent variable 'y' is expressed as a linear function of all the predictors.<input type="checkbox"/> The underlying relationship between the response and predictors is linear.<input type="checkbox"/> Also the error distribution should be normally distributed.	<ul style="list-style-type: none"><input type="checkbox"/> GLM allows us to build a linear relationship between response and predictors, even though their underlying relationship is not linear.<input type="checkbox"/> This is made possible by using link function, which links the response variable to linear model.<input type="checkbox"/> Unlike linear regression models, the error distribution of the response variable need not to be normally distributed.<input type="checkbox"/> The errors in the response variable are assumed to follow an exponential family of distribution (i.e. normal, binomial, Poisson, or gamma distributions).

swayamii 8

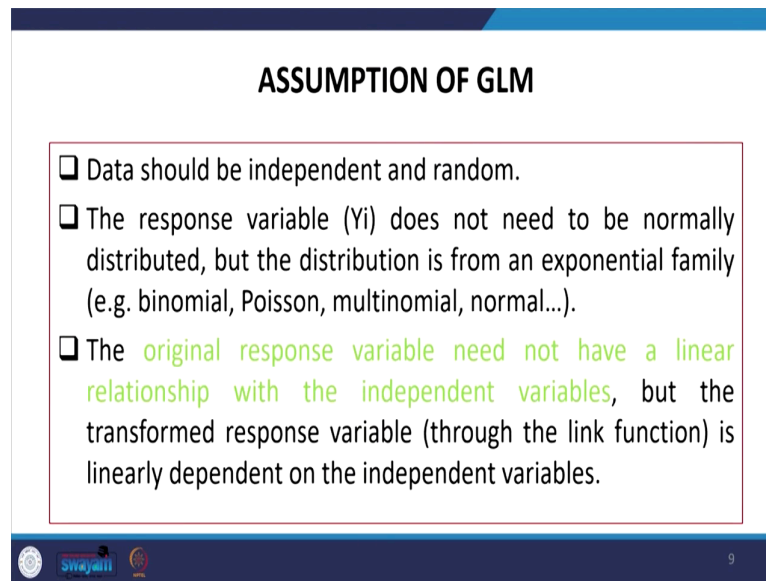
The comparison between linear model to GLM is mentioned here. In linear regression model, the dependent variable Y is expressed as a linear function of all the predictors. The underlying relationship between the response and the predictor predictors is linear. Also, the error distribution should be normally distributed.

Coming to the context of GLM, this allows us to build a linear relationship between the response and predictors, even though their underlying relationship is not linear but still it actually emerged with linear relationship with certain transformation. This is made possible by using the log link function. The way we have discussed in the previous slide which links the response variable to a linear model unlike linear regression model.

The error distribution of the response variable need not necessarily to be defined as normally distributed, the errors in the response variables are assumed to follow an exponential family of distribution.

Therefore, we say that- though the distribution are accordingly defined as may be of binomial, Poisson or gamma, but that are part of the exponential family of distribution.

(Refer Slide Time: 09:12)



The slide is titled "ASSUMPTION OF GLM" in bold black text. Below the title is a red-bordered box containing three bullet points. The first bullet point is "Data should be independent and random." The second bullet point is "The response variable (Y_i) does not need to be normally distributed, but the distribution is from an exponential family (e.g. binomial, Poisson, multinomial, normal...)." The third bullet point is "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." The text in the third bullet point is highlighted in green. At the bottom of the slide, there are logos for Swayam and a small number '9'.

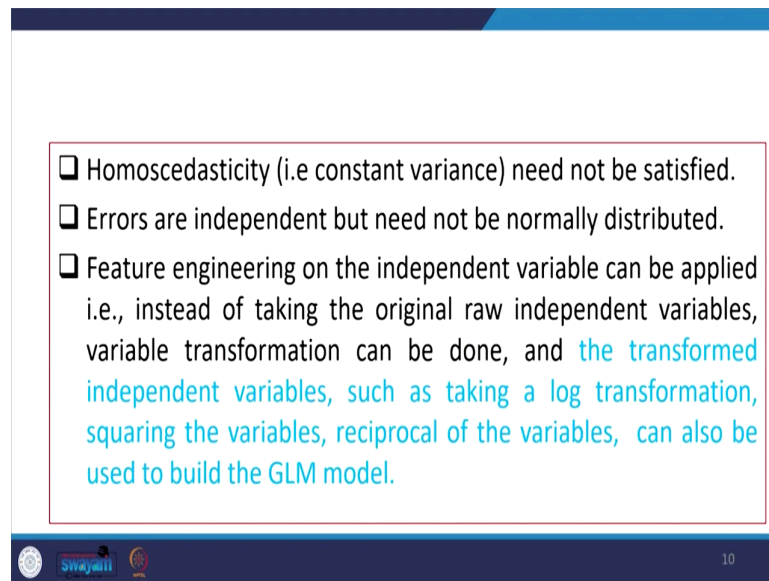
ASSUMPTION OF GLM

- ❑ Data should be independent and random.
- ❑ The response variable (Y_i) 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.

There are certain assumptions of GLM, the first one is that the data should be independent and random. The response variable does not need to be normally distributed. Response variable i.e., Y_i need not be normally distributed, but the distribution is from an exponential family nor from a normally distributed family.

So, binomial count can be poisson or multinomial or may be normal, but that should still follow a exponential family. The original response variable need not have a linear relationship with the independent variables, but the transform response variable through the link function is linearly dependent on the independent variables that we have just discussed.

(Refer Slide Time: 10:00)



- ❑ Homoscedasticity (i.e constant variance) need not be satisfied.
- ❑ Errors are independent but need not be normally distributed.
- ❑ 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.

Another assumption is that, the homoscedasticity need not be satisfied since our distribution is following a different exponential format. So, its need not required to be satisfied. So, errors are independent, but need not be normally distributed. 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 variable, such as taking a log transformation, squaring the variables, like reciprocal of the variables, can also be used to build the GLM model. So, there are various forms of transformation we do to develop a GLM model.

(Refer Slide Time: 10:52)

Various generalized least square method

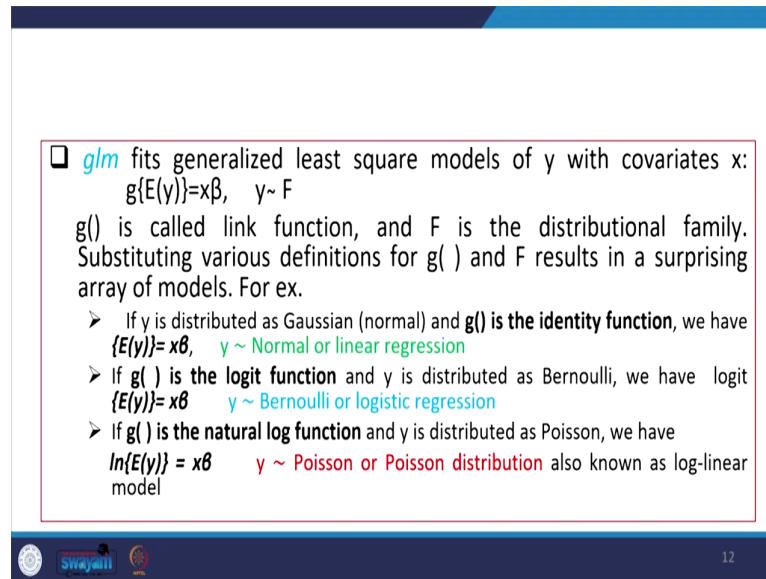
- ❑ Linear Regression, for continuous outcome with normal distribution.
 - Identity link function is used, which is the simplest link function.
 - Response is continuous
 - Predictors can be continuous or categorical, and can also be transformed.
 - Errors are distributed normally and variance is constant.
- ❑ Binary Logistic Regression, for dichotomous or binary outcomes with binomial distribution
- ❑ Poisson Regression, for count based outcomes with poisson distribution.
 - Here counts are expressed as a linear combination of the explanatory variables. Link function used here is log link function.

swayamii 11

Then there are various generalized least square methods like linear regression for continuous outcome with normal distribution. In that case, identity link function is used which is the simplest link function and the responses is continuous, predictors can be continuous or categorical or can be transformed, errors are distributed normally and variance is constant in case of linear regression. But we use binary logistic regression for dichotomous or binary outcomes with binomial distribution.

There are another type of generalized least square method. Poisson regression is used for count based outcomes with Poisson distribution. Here that counts are expressed as a linear combination of the explanatory variables. So, the link function that is used here is usually called the log link function.

(Refer Slide Time: 11:53)



□ *glm* fits generalized least square models of y with covariates x :
 $g(E(y))=x\beta$, $y \sim F$

$g(\cdot)$ is called link function, and F is the distributional family. Substituting various definitions for $g(\cdot)$ and F results in a surprising array of models. For ex.

- If y is distributed as Gaussian (normal) and $g(\cdot)$ is the identity function, we have $E(y)=x\beta$, $y \sim$ Normal or linear regression
- If $g(\cdot)$ is the logit function and y is distributed as Bernoulli, we have $E(y)=x\beta$, $y \sim$ Bernoulli or logistic regression
- If $g(\cdot)$ is the natural log function and y is distributed as Poisson, we have $\ln\{E(y)\} = x\beta$, $y \sim$ Poisson or Poisson distribution also known as log-linear model

swayamii 12

So, we follow the command that is a GLM that feeds generalized least square models of Y with covariates of x . So, further g is taken as the transformation of the expected value of Y that should follow a normal distribution with X beta.

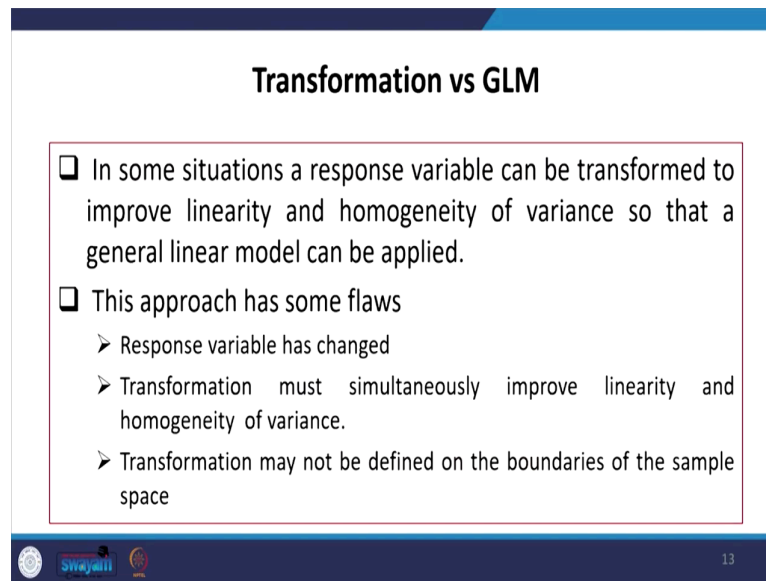
And Y follows the density function where F stands for the density function and g is called the link function because of some transformation and F is the distribution family or usually we present in the CDF type of functions. Substituting various definitions for g function and F , results in a surprising array of models.

So, if Y is distributed as Gaussian normal distribution and g is the identity function. We have the expected Y is equal to nothing but the normal distribution or normal or linear regression.

If g is a function of the variables with the transformation called g as the logit function. And Y is distributed as Bernoulli kind, we have the logit distribution that is otherwise called Bernoulli or logistic regression that follows a sigmoid kind of function where the concentration is usually at the extreme ends.

And, if g function is the natural log function. In that case, usually we go for the Poisson kind of distribution. So, Poisson distribution are for the count data. And \ln is natural log and expected value will convert the data to a linear function therefore, it is called log linear functions.

(Refer Slide Time: 13:45)



Transformation vs GLM

- ❑ In some situations a response variable can be transformed to improve linearity and homogeneity of variance so that a general linear model can be applied.
- ❑ This approach has some flaws
 - Response variable has changed
 - Transformation must simultaneously improve linearity and homogeneity of variance.
 - Transformation may not be defined on the boundaries of the sample space

swayam 13

So, now we are just comparing between transformation and GLM. In some situations, a response variable can be transformed to improve linearity and homogeneity of variance. So, that a general linear model can be applied. This approach has some flaws like- response variable has changed, transformation must simultaneously improve linearity and homogeneity of variance, transformation may not be defined on the boundaries of the sample space that has been considered. As an example, a common remedy for the variance increasing with the mean is to apply the log transformation. If variance is there, simply taking the log transformation that will reduce the variance. So, there are other ways by which we can also do some expected value of the variance, log function can also be taken.

(Refer Slide Time: 14:47)

□ For example, a common remedy for the variance increasing with the mean is to apply the log transformation e.g.

$$\log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$E\log(Y_i) = \beta_0 + \beta_1 x_i$$

This is linear model for the mean of log Y which may not be always be appropriate. E.g. if Y is income perhaps we are really interested in the mean income of population subgroups, in which case it would be better to model E(Y) using a glm $\log E(Y_i) = \beta_0 + \beta_1 x_i$

swayamii 14

Here, this is a linear model for the mean of log Y which cannot be or may not be always appropriate. For example, if Y is income, perhaps we are really interested in the mean income of population subgroups, in which case it would be better to model expected Y using the GLM. GLM of this function i.e., log expected Yi is actually converting to a linear function.

(Refer Slide Time: 15:12)

Advantage of GLM over simple linear regression

- We do not need to transform the response Y to have a normal distribution
- The choice of link is separate from the choice of random component thus have more flexibility in modeling
- If the link produces additive effects, then we do not need constant variance.
- The models are fitted via Maximum Likelihood estimation; thus optimal properties of the estimators.

swayamii 15

So, some advantages of GLM over the simple linear regression which are counted and necessary to define are- First one is that we do not need to transform the response Y to have a

normal distribution, the GLM command itself will automatically deal with this transformation.

Then, the choice of link is separate from the choice of random component and thus have more flexibility in modelling. So, the link function that we usually carry from the beginning of the command may be like log transformation or logit. At the beginning we need to think very carefully, but in this case with simple GLM command at the end we can specify the link function and derive the result.

Third, if the link produces additive effects, then we do not need constant variance. And fourth, the models are fitted via the MLE i.e., maximum likelihood estimation, thus optimal properties of the estimators are also defined.

(Refer Slide Time: 16:20)

EXAMPLE

- ❑ We fit a model based on data from a study of risk factor associated with low birth weight. (hosmer and lemshow and Sturdivant,2013)
- ❑ `glm low age lwt i.race smoke ptl ht ui, family(binomial) link(logit)`
- ❑ Following stata menu
 - Statistics > Generalized linear models > Generalized linear model (GLM)

One example we will also show you and on the basis of that you can just think of how you would be able to derive the result and be confident about your result. If we fit a model that is based on the data from a study of- risk factor associated with low birth weight as suggested in a paper by Hosmer and Lemshow and Sturdivant in 2013, where we say that the command is equal to glm then at first “dependent variable” i.e., low birth weight, and then age and then other variables, which are reference categorical variable that we can also mention.

Now, after all such variables we can give a comma and define which family they belong to. And if suppose we identify that they the data belongs to binomial family then your link

command will be of link logit link(within bracket logit) and then that will be giving you the result.

So, here are other way of deriving the command- go to statistics, then GLM generalized, linear models then generalized linear model GLM will be displayed on the screen. I will also show that to you.

(Refer Slide Time: 17:46)

Stata Command for GLM

Model of y as a function of x when y is a proportion
`glm y x, family(binomial)`

Logit model of y events occurring in 15 trials as a function of x
`glm y x, family(binomial 15) link(logit)`

Probit model of y events as a function of x using grouped data with group sizes n
`glm y x, family(binomial n) link(probit)`

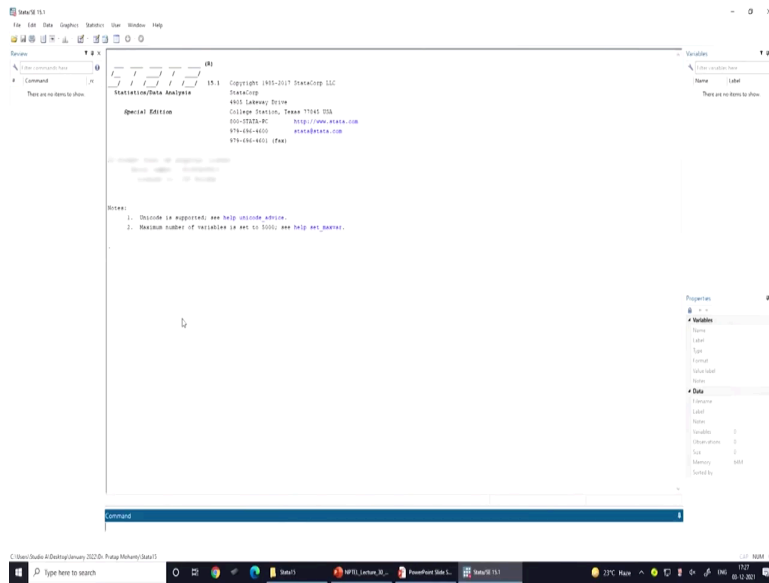
Model of discrete y with user-defined family `myfamily` and link `mylink`
`glm y x, family(myfamily) link(mylink)`

Bootstrap standard errors in a model of y as a function of x with a gamma family and log link
`glm y x, family(gamma) link(log) vce(bootstrap)`

17

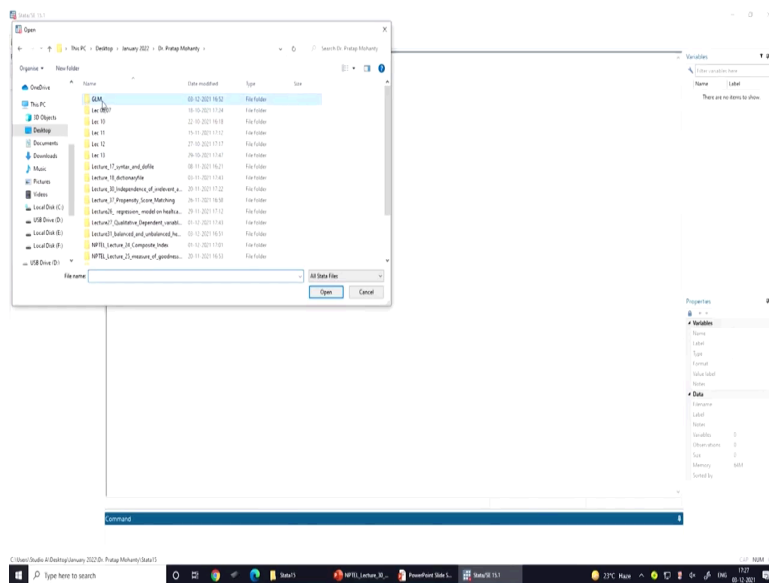
Stata command for GLM- there are different command schedule. I am just going to show you that how you can operate just for your reference.

(Refer Slide Time: 17:56)

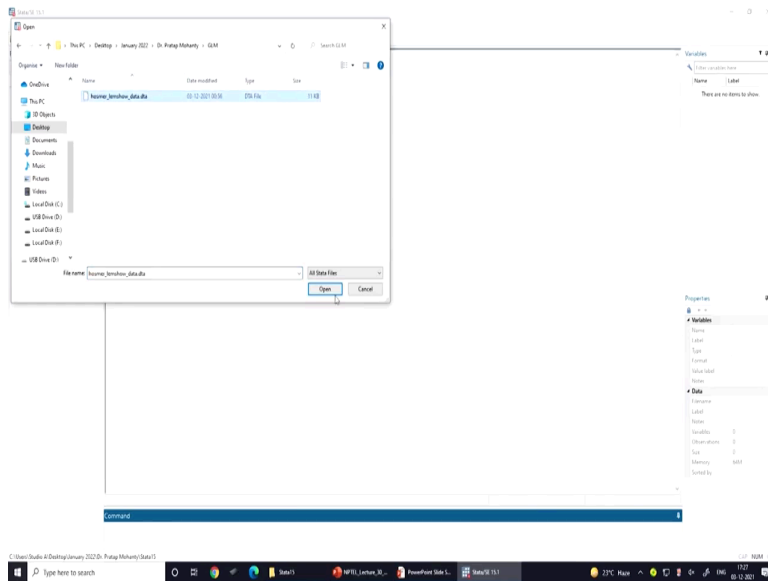


Here is the data. Now we want to operate it once GLM is here.

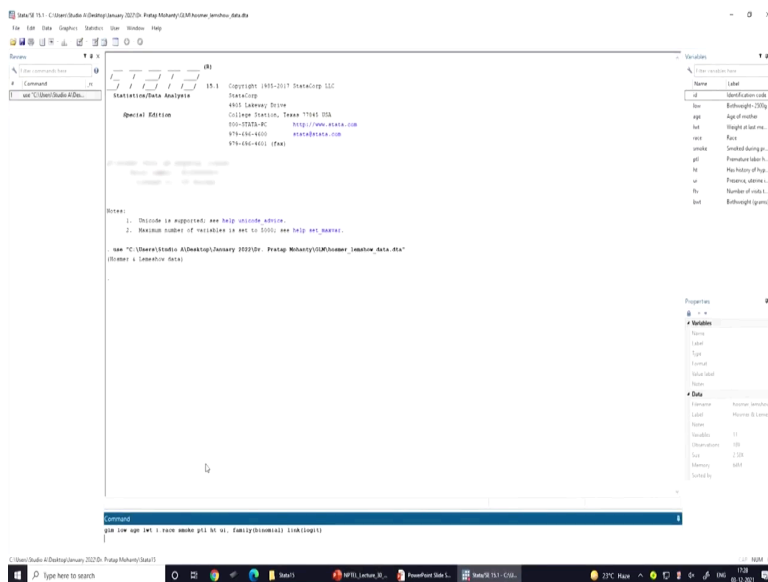
(Refer Slide Time: 18:05)



(Refer Slide Time: 18:06)

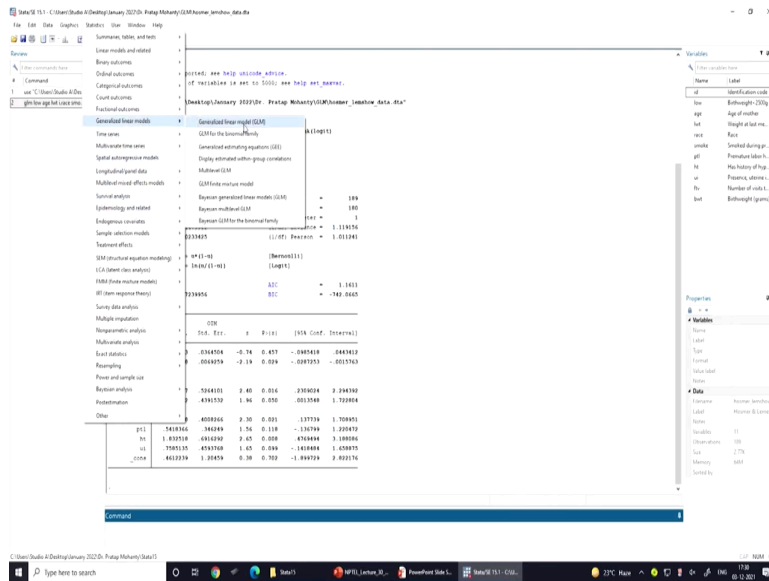


(Refer Slide Time: 18:09)



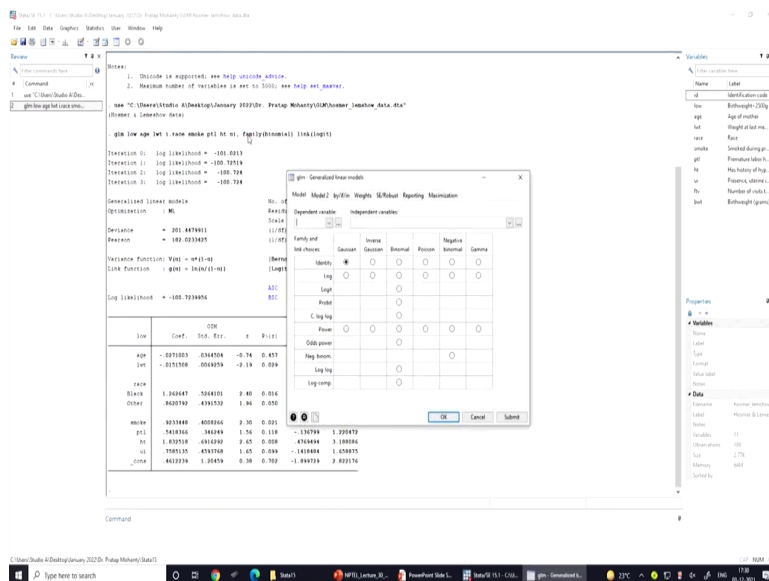
So, the sample data we are just operating here for your reference, we have all those data and we will also open the do file simultaneously or we can just directly run as per the command we have given here is the command section. So, we simply copy from do file and you can run for your own practice.

(Refer Slide Time: 20:17)



You can go to statistics then it has what is called generalized linear models.

(Refer Slide Time: 20:29)



Then here is your GLM now you can specify each of it, i.e., the dependent and independent variable and with that it is possible for you and some family name you can give it. So, with the family name like as I already discussed whether it is a Gaussian type, binomial, Poisson, gamma etc. and your result will be displayed on your screen the way we have shown it to you.

So, these is all in this lecture. I think at this moment we will come up with further details in our next episode and with this we are going to stop here. And I look forward to your comments or queries and we will be happy to address it.

Thank you.