

Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 57
Fitting Models to Data

Hello everybody, this is Rudra Pradhan here, welcome to Engineering Econometrics. Today we will start with a new concept that is on count data and discrete modeling. We have already discussed various types of econometric modeling starting with linearity structure, nonlinearity structure, cross sectional kind of modeling, time series kind of you know modeling. And within time series modeling we have discuss various types of you know structures starting with linearity to non-linearity, simple structure to volatility modeling.

And we have also covered the concept called as a panel data modeling where we use actually time series structure and cross sectional structure. In that too we have covered couple of items under the panel data modeling like pool data structure then a fixed effect models, random effects model, generalized method some methods of moments. Likewise we have you know discuss you know various aspects of you know engineering economics starting with simple to complex. And with the lots of variety, lots of flexibility, depending upon the type of you know engineering problems and the kind of you know requirement, maybe organizational requirement, maybe corporate requirement, maybe industrial requirements.

So, technically we have highlighted couple of you know engineering economics tools through which we can solve some of the problems depending upon the particular requirements, maybe the industrial aspects or kind of you know environmental aspects. You know typically you know it depends upon the type of you know engineering problems. Whatever we have discussed already basically the choice of the technique and the kind of you know models which you like to use or we like to deploy to solve some of the engineering problems.

One thing is very you know clear that is very important; so, for as a choice of a modeling is concerned to select a particular model for solving a particular engineering problem. So, the idea is that you know you should not the kind of you know you know data

structure. So; that means, over and above what above whatever types of you know problems we are going to discuss and the kind of you know models which you like to pick up exclusively depends upon the types of you know data sometimes.

For instance, we have in the first instance we have 3 types of you know structure that is what called as you know a cross sectional structure, time series structure and panel structure. So, corresponding to these structures so, we have called as you know cross sectional data, we have time series data, we have pool data and we have panel data.

While solving any problems through any models of course, we have to pick up a problems depending upon the problems requirement and the kind of you know engineering requirement. But, still sometimes the choice of a particular model is data driven if for instance; if the time series data is not available so, you cannot use time series modeling. Similarly, if the data is not with respect to both time series and cross sectional type you cannot deploy pool data panel data modeling.

So; that means, what I like to you know bring here it is the type of data sometimes very vital that too choice of a particular technique and so, far as the particular problem is concerned. The problem is such a nature where you have to strictly follow a particular technique no other techniques can help to solve that problem.

So, one instance the choice of technique that too with respect to problem specific and on the other side it is the choice of technique with respect to data structure. So, where your data is actually behaving time series then you can simply deploy you know time series modeling. If the data is a panel kind of you know structures you can you know apply panel data modeling or if the data is simply cross sectional then you can you know you simple cross sectional modeling.

So, likewise we have you know lots of you know choice, lots of variety. You can pick up any technique corresponding to the data structures and corresponding to the problem requirement and then as usual the procedures are you know more or less same. Somewhere we use complex kind of you know tool, some somewhere we use actually simple kind of you know structure, ultimately the processing is this more or less actually similar.

So, we have to just you know pick up the problem, set the objective, build the hypothesis, then you know build the model. And then you know with the help of the data and the choice of techniques you have to estimate the model and after receiving the estimated you know outcome. So, we have to go through lots of you know diagnostics.

And then we come with a kind of you know model we should be very solid one with respect to the kind of you know diagnostics and sometimes we use robustness. Then we come with a kind of you know structure where the model will be perfectly feed to analyze the particular you know engineering problem.

On the top of these we like to bring a situation now which is slightly different compare to whatever we have discussed. So, for as types of you know models that too with respect to the problems and the kind of you know that data is concerned. In a kind of you know analytics framework we may have a structure of you know statistical kind and we have a structure like you know mathematical kind.

So, for a mathematical optimization is concerns so, there is a concept called as you know integer programming. If there is no such restriction then you can optimize a particular problem and look for the values of the decision variable. And then that values of the decision variable can you know solve the problems as per the decision making is concerned. But, when you put the restriction that you know the values of the decision variable will be integer type then the problem is not so, simple.

So, you to continuously monitor till you get the values of the decision variable which is integer type and that values of the decision variable can help to solve the engineering problems and as for the decision making requirement. In the econometrics we have the similar kind of you know flow. Till now whatever we have discussed, whatever models we have discussed, whatever problems we have connected to simplify or to analyze to who have no such restrictions about the kind of you know you know integer types of you know situation.

For instance, all these you know discussions or all this problems which you have highlighted is a data driven where we have no such restriction about the types of you know data. Whether it is a time series, weather it is a cross sectional and whether it is a kind of you know pool or you know penal. But, bringing the structure of you know time series, cross sectional pool and panel is a one kind of you know flow. The other flow is

that you know the kind of you know information within the kind of you know particular structure where the information may be numeric in nature or it may be qualitative in nature.

So, so, we have already discussed couple of you know problems or couple of models relating to the type of you know data that too information behind the data structure. We have no issue about the time series or cross sectional or you know pool or panel but, what is more important in the second structure is the type of you know information relating to the particular you know variable. So, that information may be typically you know qualitative in natures or sometimes it may be completely numeric in nature.

If the particular informations for a particular variable is completely numerical so, most of the problems is very straight forwards and you can analyze without any difficulty. Of course, sometimes we can do the transformation do the restructuring. So, it is as per the requirement of the problems or to bring the best. But, here the situation is a different, sometimes the information by default will be qualitative in nature.

Of course, we have already discussed couple of models you know in the similar angle for instance, the classic example is the dummy modeling where we have use we use actually linear probability models or otherwise it is called as a binary choice models. Then we have discussed logit model, we have discussed probit model where the kind of you know structured is actually means is dummy type. That means, the variables which is actually a categorical in nature or sometimes we called as you know binary in nature sometimes we call called as you know you know qualitative in nature. So, the we have already discussed you know several types of you know problems and the type of you know models.

For instance, if you start with the binary choice model then there are 2 things you know means there are 2 requirements all together. First requirement is the independent variable which is the qualitative in nature and that too information must be lie between 0 to 1. And we have no such restriction on the independent variables that is what the model is about the in linear probability model or binary choice model. And now it is up to you where we can actually use this models and you know what are the requirements for this and what type of you know engineering problems you to apply.

Similarly, in the case of logit model, it is actually slightly similar but, little bit different because, both logit model and probit model are non-linear in character compared to the linear probability model. So, here we have the similar kind of you know structure. The dependent variable will be a dummy type that is categorical in nature or qualitative in nature where we have no such restriction on independent variables.

However, if we compare with linear probability model or binary choice model in you know in the case of you know logit model and probit model. So, it is the probability value which will bring the functional form and compared to linear probability models where the p value will be that is the probability value which will be in between 0 to 1, that is why it is called as you know binary choice model. But, in the case of logit model and probit model there is no such restriction. Of course, the information will be qualitative like the binary choice model. And again exclusively represented by the probability.

And what is actually happening here the probability value will not you know within the 2 extremes where the binary choice model you know you know express. It is the case where the probability value which will be in between 0 to 1 and that is what the differentiating point between logit and probit that too with you know binary choice model. So; that means, the probability value which will be in between 0 to 1.

So; that means, here technically if it is in between 0 to 1 then technically the values of the probability will be fractional type like $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$ and so, on. Because, probability value cannot actually exceed 1 and cannot be against you know in between you know say means basically it will be in between 0 to 1. So, it will not be this side or it will not be that side. So, it will be in a kind of you know structure like this. So, you have to bring that structure and then apply these models. So, here there is no such restriction but, only restriction is p value will be in between 0 to 1. So; that means, completed the information must be fractional type.

Now, we are discussing 2 different situations, in one situation we need dependent variable, qualitative in nature that too the classification will be 0 to 1. The corresponding problems may be like you know yes, no, male, female something like that. In other situations it is in between actually 0 to 1 where we can deploy logit model and probit model.

And the information for this you know logit model and the probit model that too for the dependent variable is again qualitative in nature and not too extremes between 0 and 1, it is you know between 0 to 1 only. So, like you know 1 by 2, 1 by 3, 1 by 4 and so, on. So, that is how you know so, far as you know data structuring concerns we use group data in the case of you know logit model and probit model and we use individual data in the case of you know linear probability model.

And by the way we have already discussed you know all these models with various examples and the kind of you know kind of you know problems and structure. And we are in a comfortable positions for know what is exactly binary choice model, what is exactly logit model, what is exactly probit model and where you have to apply, how to apply, how to bring the estimated output and how to interpret and then how we can actually use as a decision making tool.

Now, coming to this you know lectures where we have similar kind of you know flow and the flow is actually with respect to count data and discrete modeling. It is it is in the similar kind of you know basket little bit you know means we have a let us say we have a 3 difference structure altogether simple.

We know structures where any kind of you know modelling can be used without any restriction on data. You may use time series data, you may use data, you can use pool data, you can use panel data. But, second level of you know modeling structure will be where we have some kind of you know restriction that too either with the dummy in dependent variables or with respect to independent variable or with respect to both dependent variable and independent variable.

If it is with respect to only independent variable that is called as a dummy independent modeling and at least one of the independent variable will be qualitative in nature and where we have no such restriction whether the kind of you know information will be 0, 1 or 1 by 2, 1 by 3 something like that. But, it will be somehow you know classified in a some you know structure so that you know it can be apply as for the modeling requirement.

But, another kind of you know structure is dummy dependent modeling where we have again to different you know setups. In the first set ups the dependent variable will be qualitative and that too information will be in between 0 to 1 only, either 0 or 1. And the

structure will be again the information relating to dependent variable will be qualitative in nature and that too the information will not 0 to 1, it will be in between 0 to 1 so; that means, exclusively fractional type.

Now, the third layers of the models will be slightly complex and you know very exceptional kind of you know situations. Against we may not have issue about the kind of you know dependent variable and independent variable. But, it is the variables that too what information will be you know information will be the differentiating factors; that means, we have discussed the dependent variable structure, independent variable structure that too dummy types and the kind of you know fractional type and 0, 1 type.

Now, this is the kind of you know model where we declare as you know count data and discrete modeling where the variable informations will be exclusively integer type. So, that is how the counting is there so; that means, it is again very tough class kind of you know data structures. And very restricted kind of you know environment where we can apply this kind of you know modeling.

Because, this kind of you know modeling very strictly apply to those situations where the data structure is actually like you know count data and kind of you know discrete data. So, obviously, we like to know what is exactly the concept count data and the discrete data then we will follow the count data and discrete data modeling. So; that means, this particular lecture exclusively meant for regression modeling with count data only count data.

And you know the kind of you know discrete data, which is slightly different you know from the other modeling clusters, where we have use without any restriction. And again this second cluster there is some restriction with dependent variable and independent variable, again 2 layers of you know restriction to the dependent variable. That is what the you know different kind of you know engineering econometrics classifications so, for as you know econometrics models are concerned.

So, now this is kind of you know situations where we have to understand first count data that is and the discrete data. And then we like to understand or you like to know what kind of you know models have to bring in those situation. That is very important here so; that means, the problem understanding the type of you know data and the selection of

you know models again. So, it is basically regression with modeling with the count data or in other words called as a count data modeling.



So, let us start with you know understanding the count data, then we will go through the various models relating to count data and then we connect some of the examples related to count data modeling where we can apply count data modeling because, the data specification is count data type. So; that means, it is a very data you know data driven techniques altogether.

So, problem is very specialize in character it may be any engineering problems. But, it is it is completely different than other types of you know econometric modeling which we have already discussed. And most important is that you know we have know you know big issue relating to this you know time series, cross sectional pool and penal. But, here one of the biggest you know issue is the kind of you know information lying in the data that too you know the informations you know for a particular you know variables. So, that is what the big deal.

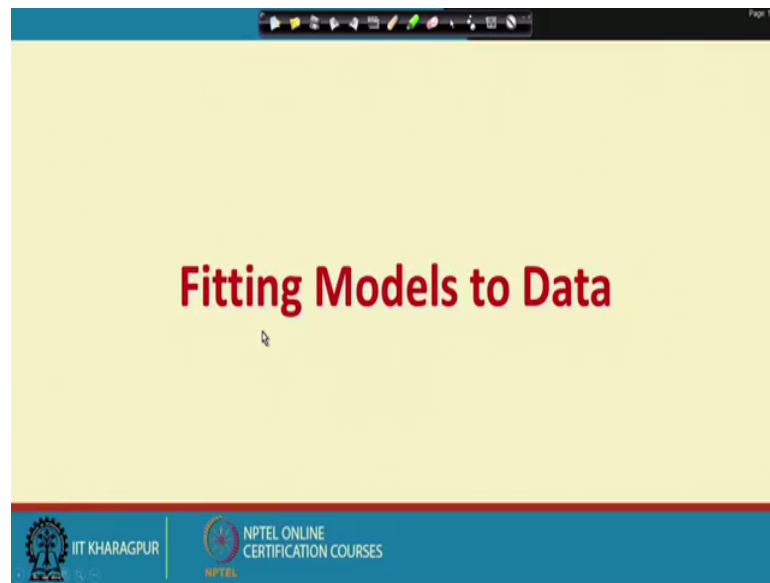
So, let us let us understand what is that deal and then we will go to this you know a model and then we connect with the some of the application relating to this you know count data modeling.

(Refer Slide Time: 22:49)

Course Contents	
Weeks	Lecture Names
Week 1	: Introduction to Engineering Econometrics
Week 2	: Exploring Data and Basic Econometrics on Spreadsheets
Week 3	: Descriptive Econometrics
Week 4	: Linear Regression Modelling
Week 5	: Modelling Diagnostics 1
Week 6	: Modelling Diagnostics 2
Week 7	: Non-linear Regression Modelling
Week 8	: Time Series Modelling 1
Week 9	: Time Series Modelling 2
Week 10	: Panel Data Modelling
Week 11	: Count Data and Discrete Modelling
Week 12	: Duration Modelling

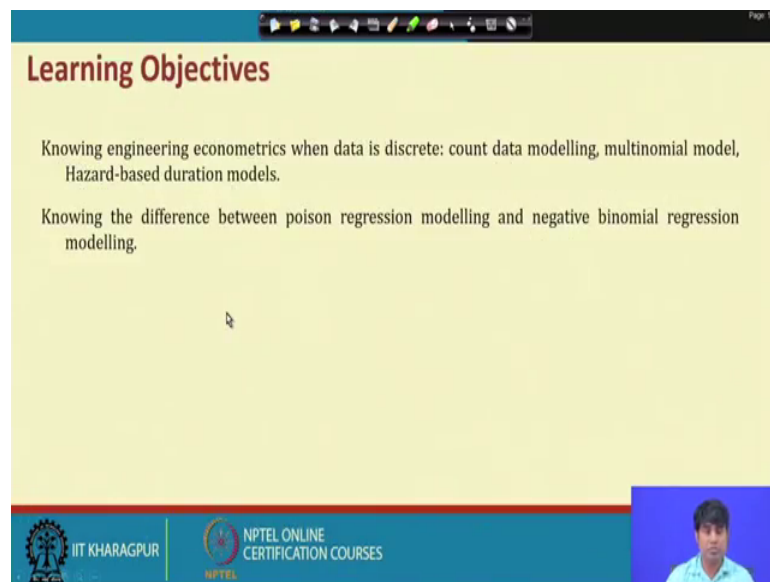
 IIT KHARAGPUR  NPTEL ONLINE CERTIFICATION COURSES 2

(Refer Slide Time: 22:51)



So, the first understanding is the you know about the you know so; that means, technically.

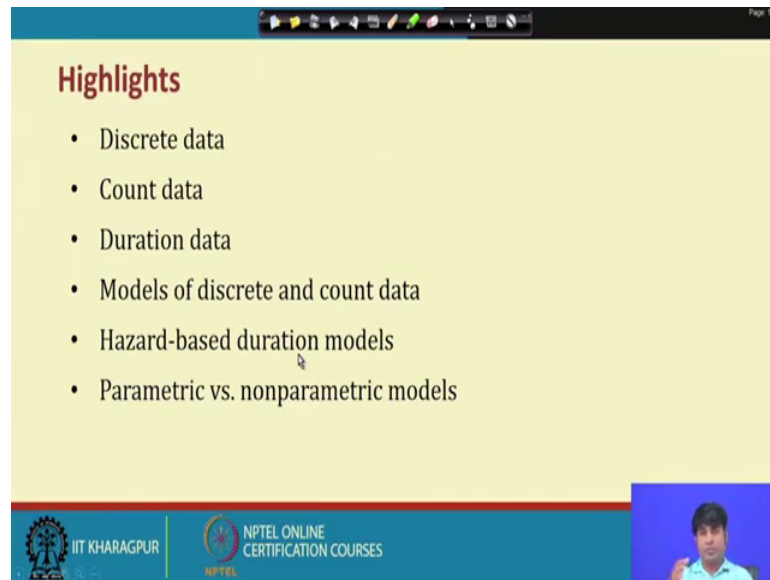
(Refer Slide Time: 22:55)



We called as you know regression modeling with count data or fitting models with count data we can do whatever you like. So, we have to objectives here and the basic objective is to understand the count data and the kind of you know discrete data. And then we second objective it is to know the various models relating to count data and you know discrete data. And then the third objective is what are the areas or you know: what are the

application through which we can use count data modeling and provided the data should be count data in character right. So, these are the 3 objective through which you can actually.

(Refer Slide Time: 23:35)



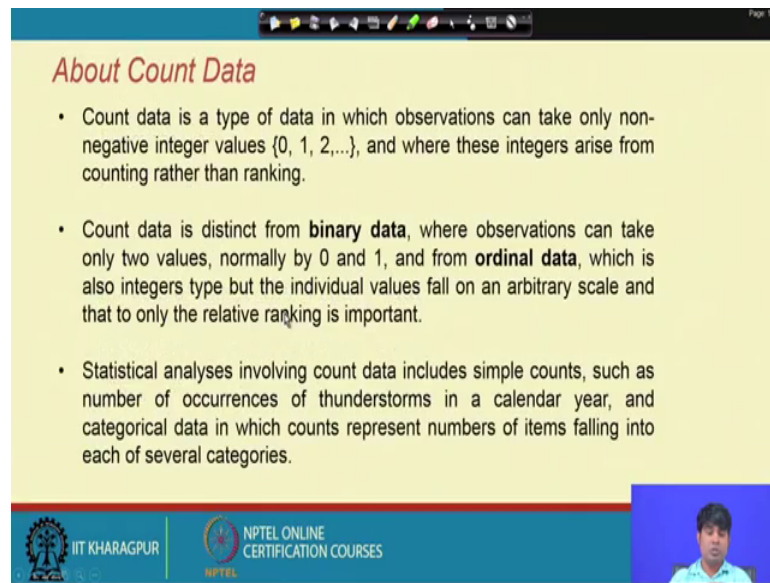
Highlights

- Discrete data
- Count data
- Duration data
- Models of discrete and count data
- Hazard-based duration models
- Parametric vs. nonparametric models

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Start the game about the count data modeling or fitting the model with count data. So; obviously, you know some of this highlights will be to understand the discrete data, count data and there is a concept called as a duration data. And models of discrete and count data hazard based duration models and then the structure about the parametric and nonparametric you know regression models.

(Refer Slide Time: 24:04)



About Count Data

- Count data is a type of data in which observations can take only non-negative integer values $\{0, 1, 2, \dots\}$, and where these integers arise from counting rather than ranking.
- Count data is distinct from **binary data**, where observations can take only two values, normally by 0 and 1, and from **ordinal data**, which is also integers type but the individual values fall on an arbitrary scale and that to only the relative ranking is important.
- Statistical analyses involving count data includes simple counts, such as number of occurrences of thunderstorms in a calendar year, and categorical data in which counts represent numbers of items falling into each of several categories.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Page 11

So, now we start with first to understanding the count data. So, already I have you know you know highlighted this concept. So, once again you know to bring the notice about the count data that you know count data is a type of you know data which is the a special prize in character in which observations can take only non negative integer values, that is what the actually biggest you know restriction non negative integer value. So, it is the integer values and again it is called as you know non negativity means, exclusively in a kind of you know optimization environment you know in a kind of you know analytic frameworks. So, where it is a we called as you know integer programming.

So, in the integer programming problems we have 2 different requirements and you know restriction the values of the decision variable will be positive in natures that is what be non negativity. And the values of the decision variable will be integer types; that means, it is it will be only 0, 1, 2, 3, 4 and something like that. The fractional values are not actually allowed there. So, same concept we are just bringing here so; that means, there actually the problem is there we are optimizing the situation and looking for the values of the decision variable. Where the values will be integer type and non negative in character, then you can apply and analyze the situation.

But, here the situation is slightly different and what is happening here? We are going in a kind of you know data driven to model you know in a kind of integer programming it is model to values of the decision variable. So, here I like to start with you know the

concept called as you know data and then feed the model and then you know use this model for you know problem requirement.

So, basically count data is a type of data in which observations can take only non negative integer values like y 0, 1, 2 and so, on. And these integers you know arise from counting rather than ranking that is very important. So, it is only counting so, number of frauds, number of deaths ah, number of you know females, number of you know kind of you know accidents, number of injuries. So, these are the you know kind of you know items where we bring the concept called as you know count data.

It is not the question of ranking. Ranking means you know it is the degree of moment for instance you know very good best like this kind of you know things. It is a slightly you know different with respect to counting. So, it is call you know exactly we cannot call it is a qualitative in nature. But, the structuring of the data is slightly different compared to the kind of ranking.

So, ranking by default will make you know difference about the data points or the understanding of the variable. So, for as information is concerned for instance, good very good person so, there is a difference actually that is the degree of difference and exactly to quantify the difference is very difficult. But, we can recognize that there is a difference but, what is the beauty of the count data that you know it can exactly quantify the particular you know difference.

So, it is not ranking but, it is the kind of you know clustering where you know we like to counts you know in numbers only and that numbers should be integer types. When we count numbers it cannot be negative. Of course you know the expression maybe starts with the negative but, reporting will be reporting will be in numbers only and that too with it you know 2 restrictions, one restriction is the values will be positive in character and the second one is the it is exclusively integer type. So, no fractional will be allowed here.

So, for instance know why not you know the obvious question is that why not. So, for instance let us say you know number of deaths is a kind of you know example. So, half of the death cannot be you know means it has you know meaning at alls so, half death half accident. So, these are all not actually you know meaningful kind of you know information or meaningful kind of you know presentation.

So, if you instead of saying you know half injury so, you can better to say number of injury, instead of in half death you can say that you know number of deaths. So, half death there is no such you know meaningful representation. So; that means, technically you need lots of adjustment and understanding, then structuring, just to bring the kind of you know information that too declared as the count data. And then use this count data for any kind of you know modeling and that too what we called as you know count data modeling and discrete data modeling.

As per the you know discussion what we have already highlighted count data as is slightly distinct from you know binary data like which I have already highlighted in the case of you know linear probability model and binary choice model. The there the requirement is actually the information should be qualitative in nature but, we transfer into some coding and that coding must be 0, 1 interval only either 0 or 1 so, it is not in between.

So; that means, count data exclusively different from binary data first instance and second again count data is different from the ordinal data where we need some kind of you know ranking ok. So, like you know good, better, best something like that you know high, low, very high, very low, these are all different kind of labeling and through which ranking and some kind of you know skilling will do. But, count data is you know you know is a something but, it is a very dangerous kind of you know situation where you have to deploy or you know use you know then that too actually as per the requirement. So, statistically and then analyze the situation.

So, so, what is actually important here just to understand the count data and being the difference of count data with respect to various other types of you know data starting with the categorical data, qualitative data. And of course, here there is no such you know issue with respect to time series, cross sectional and penal data. But, it is the information which can make the difference and the too the difference with respect to models and the kind of Y analysis the kind of you know structuring. So, that is very important here so, in.

So, we will stop here, in the next lectures we will start from here to connect various models relating to count data and you know discrete data.

Thank you very much, have a nice day.