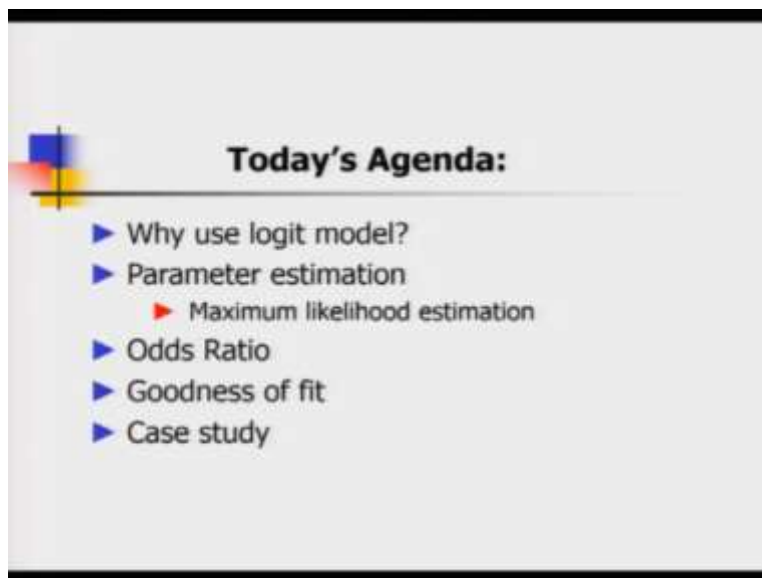


**Applied Statistics and Econometrics**  
**Professor Deep Mukherjee**  
**Department of Economic Sciences**  
**Indian Institute of Technology, Kanpur**  
**Lecture 36**  
**Dummy Dependent Variable Models Part II**

Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So, today we are going to continue our discussion on discrete choice models or qualitative dependent dummy variables. So, here in this lecture we are going to talk about logit model.

(Refer Slide Time: 00.36)

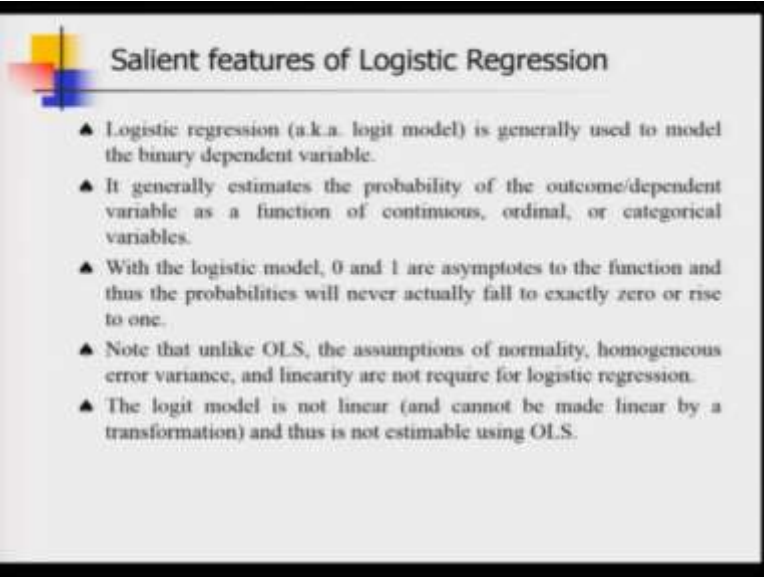


So, let us have today's agenda items. So, we are going to motivate you why we require logit model. So in a nutshell I am going to talk little bit about what we have discussed in the last lecture, so it will be kind of a recap. Then we are going to talk about the formation of logit model and then parameter estimation of logic model. So in this lecture I am going to introduce a new estimation method as against two ols that we have learnt.

So this one is called Maximum likelihood estimation. Then we are going to revisit the concept of odds ratio and see how it is linked to logit model and if we estimate any regression model, we are always interested in goodness of fit. So we are going to talk briefly about goodness of fit of a logit model and then finally we will wrap up by talking about case study.

So, last time I have shown you how the linear probability model which is basically the simple linear regression model that we have studied earlier, if you apply that to this kind of discrete valued dependent variable case, then you are going to get faulty results. So what are the solutions? In the last lecture we have discussed very briefly about one solution and that is called probit model and in today's lecture we are going to discuss the other case which is the logit model.

(Refer Slide Time: 02:07)



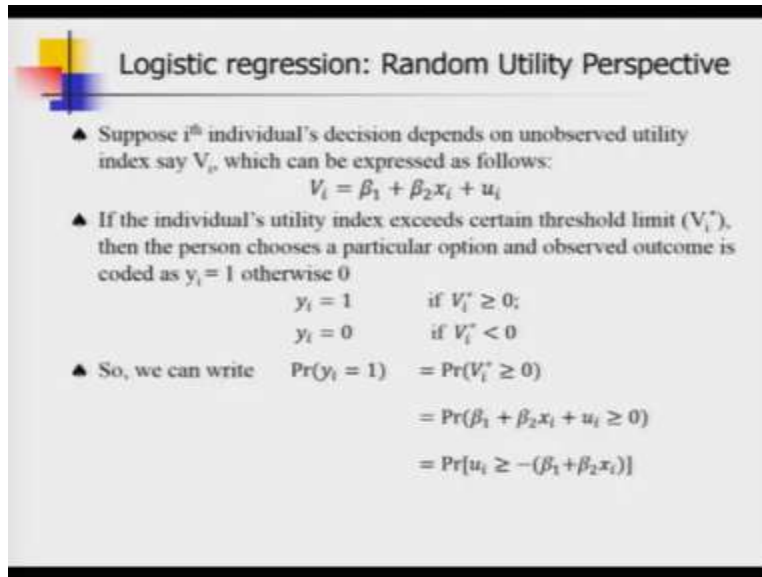
**Salient features of Logistic Regression**

- ▲ Logistic regression (a.k.a. logit model) is generally used to model the binary dependent variable.
- ▲ It generally estimates the probability of the outcome/dependent variable as a function of continuous, ordinal, or categorical variables.
- ▲ With the logistic model, 0 and 1 are asymptotes to the function and thus the probabilities will never actually fall to exactly zero or rise to one.
- ▲ Note that unlike OLS, the assumptions of normality, homogeneous error variance, and linearity are not required for logistic regression.
- ▲ The logit model is not linear (and cannot be made linear by a transformation) and thus is not estimable using OLS.

So, a logit model also estimates the probability of outcome or dependent variable as a function of continuous ordinal or categorical variables like probit model does. So, likewise logistic model always reports fitted probability values which are between 0 and 1 all the times and note that unlike OLS, the assumptions of normality, homogeneous error variance and linearity are not required for logistic regression and interestingly also note that the problem we are dealing, the regression problem we are dealing here is non-linear in nature.

So, OLS is out of question. So, we have to use you know a new estimation method which can handle non-linearity.

(Refer Slide Time: 02:55)



**Logistic regression: Random Utility Perspective**

- ▲ Suppose  $i^{\text{th}}$  individual's decision depends on unobserved utility index say  $V_i$ , which can be expressed as follows:  
$$V_i = \beta_1 + \beta_2 x_i + u_i$$
- ▲ If the individual's utility index exceeds certain threshold limit ( $V_i^*$ ), then the person chooses a particular option and observed outcome is coded as  $y_i = 1$  otherwise 0  
$$y_i = 1 \quad \text{if } V_i^* \geq 0;$$
$$y_i = 0 \quad \text{if } V_i^* < 0$$
- ▲ So, we can write  $\Pr(y_i = 1) = \Pr(V_i^* \geq 0)$   
$$= \Pr(\beta_1 + \beta_2 x_i + u_i \geq 0)$$
$$= \Pr\{u_i \geq -(\beta_1 + \beta_2 x_i)\}$$

So, now I am going to talk about the formation of logistic regression model and this time I am going to bring the microeconomic perspective of logistic regression. So there is a model called Random Utility Model. What is utility? So, those who are from economics background, utility is not a new concept to them.

We all know about utility functions and various functional forms and all but those who are taking this course or listening to these lectures but are not from the economics background, for them let me give a very simple definition of utility. So utility is basically some kind of pleasure or satisfaction out of consumption of any particular good or service or if you know there is an event happening around you, then is there any satisfaction from that event that you are drawing. So this is basically in a nutshell utility is all about, it is basically psychological state of mind, okay.

So, with this brief introduction to utility, let us go back to the model. So, now here you see we have unobserved utility index and of course this is not at all irrational to assume because utilities are basically state of mind, so they are unobserved and let us denote this unobserved utility level of  $i^{\text{th}}$  individual by  $v_i$  and let us assume that utility depends on some explanatory variables and also, we can write a regression equation for that.

So, here I am assuming that there is only one explanatory variable  $x$  which is determining the values of utility and that is why you know we have this simple two variable regression model. Of

course  $u_i$  is the stochastic random error term. Now we say that if an individual's utility exceeds certain threshold limit, say  $v^*_{i}$  is that threshold value, then the person chooses a particular option and observed outcome is coded as  $y_i$  equal to 1 otherwise it is 0.

So, let me explain this thing that I have just spoken about, so here let us give an example, suppose we are talking about person deciding whether to purchase a health insurance coverage or not. So if the person ultimately purchases health insurance coverage, we say that that is a positive outcome and we want to code that as  $y$  equals to 1 and if the person does not purchase the health insurance coverage then we say that we are going to code this outcome negative outcome by 0,  $y$  will take value 0.

So  $y$  is basically a dichotomous variable, so the explanatory variable takes only two values 0 and 1 and these two are two different choices to be made by the consumer. So now how will person decide whether to go for health insurance or not? So there are several factors which may be determining a person's decision making process, it could be his age, his education level, his income level, whether he has children at home or not there could be many-many explanatory variables but the point is that after every explanatory variable's role taken care of, there is some threshold utility level that consumer is looking at.  $S$

o, this is called reservation utility. So if buying this health insurance coverage gives him or her a utility level which is higher than this reservation utility, then he or she will definitely purchase this insurance product. So, that is the way a consumers psyche actually works. So you know this reservation utilities actually is denoted by this threshold utility level,  $v^*_{i}$ .

So now let us come back to the model here, so here you know this is mathematically represented as  $y_i$  equals to 1 if  $v^*_{i}$  is greater than or equal to 0 and  $y_i$  is equal to 0 if  $v^*_{i}$  is less than 0. So that I have already explained you, so here I am assuming that this reservation utility level is 0. So, you can say that if the net benefit of that health insurance cover to the individual is positive then there is positive utility from having one unit of that product and that is above the reservation utility which is benchmark zero.

So the person will go and purchase ah the health insurance product, otherwise he will not purchase it. So now we are interested in the probability of this positive outcome that the person

has chosen to purchase an insurance cover or it can be any other event. So why we are interested in probability?


I have already explained in the last lecture that in the discrete choice model cases when we are dealing with dichotomous explanatory variable, we actually aim to model probability and not the values 1 and 0, that we observe for the dependent variable or the outcome variable. So now let us translate this probability of  $y_i$  equals to 1 into the unobserved utility term. So this probability is equal to this probability that  $v^*_{i}$  is greater than or equal to 0.

So, now we have this expression for  $v_i$ , so we plug that expression here in place of  $v^*_{i}$  and so we can see that if  $\beta_1 + \beta_2 \text{ times } x_i + u_i$  gives a value which is greater than or equal to 0, then the person is going to go for this option, say purchase of health insurance coverage. So now note that this can be rewritten in terms of this stochastic random error variable that you have because you see ultimately in this expression  $x_i$  is basically a deterministic variable.

So the stochasticity or randomness will come from only the  $u$  variable here. So that is why the probability is given for the specific value of  $u_i$ . So ultimately, we translate the probability of observing that one individual has chosen one particular option into a probability involving the stochastic disturbance term  $u_i$ .

Now you know of course this probability will come from some probability distribution and we have flexibility to choose any probability distribution and last time if you remember, we said that in the case of probit, we have chosen for normal distribution. So here you know we are going to choose another distribution and that will actually lead us to the logit model.


(Refer Slide Time: 10:43)



### Logistic regression setup (contd.)

- Assume the probability distribution is symmetric around zero. So, we can write
$$\Pr\{u_i \geq -(\beta_1 + \beta_2 x_i)\} = \Pr\{u_i \leq (\beta_1 + \beta_2 x_i)\}$$
- Therefore
$$P_i = \Pr(y_i = 1) = \Pr\{u_i \leq (\beta_1 + \beta_2 x_i)\}$$
- We assume that  $u_i$  follows logistic distribution
- So,  $\Pr(y_i = 1)$  can be given as
$$P_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x_i)}} = \frac{1}{1 + e^{-x_i}}$$
- Similarly  $\Pr(y_i = 0)$  is defined as
$$1 - P_i = \frac{1}{1 + e^{x_i}}$$

Note: CDF of logistic distribution has a specific functional form, defined in terms of exponential function.



### Logistic regression: Random Utility Perspective

- Suppose  $i^{\text{th}}$  individual's decision depends on unobserved utility index say  $V_i$ , which can be expressed as follows:
$$V_i = \beta_1 + \beta_2 x_i + u_i$$
- If the individual's utility index exceeds certain threshold limit ( $V_i^*$ ), then the person chooses a particular option and observed outcome is coded as  $y_i = 1$  otherwise 0
$$y_i = 1 \quad \text{if } V_i^* \geq 0;$$
$$y_i = 0 \quad \text{if } V_i^* < 0$$
- So, we can write
$$\Pr(y_i = 1) = \Pr(V_i^* \geq 0)$$
$$= \Pr(\beta_1 + \beta_2 x_i + u_i \geq 0)$$
$$= \Pr\{u_i \geq -(\beta_1 + \beta_2 x_i)\}$$

So, we have to make an assumption that the probability distribution is symmetric around 0, so we can write the following statement under the bullet point 1 and then from that statement, we can write another statement that explains how we are going to calculate theoretically probability of  $y_i$  equals to 1. And now let us assume that my stochastic disturbance term  $u_i$  follows a logistic distribution. So it is a new distribution that we have introduced in the course.

I am not going to talk a lot about the shape or the moments of the logistic distribution because that is not the objective here. The objective is that this logistic distribution is very handy to

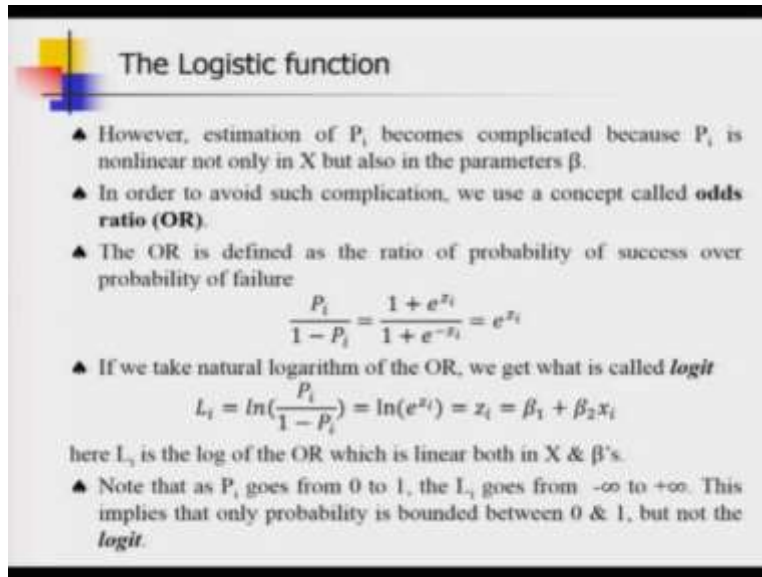
model this kind of dichotomous  $y$ . So that is why we are bringing that here and we are going to only see that if we assume logistic distribution, then what will happen?

So here you see that if I am modeling probability of  $y_i$  equals to 1, as you know logistic distribution then I can actually write this expression  $\pi_i$  equals to 1 divided by 1 plus  $e$  to the power minus  $\beta_1$  plus  $\beta_2$  times  $x_i$ . So basically here this is  $\pi_i$  is basically the probability of the positive outcome, that is  $y_i$  equals to 1 and note that I can assume that  $\beta_1$  plus  $\beta_2$  times  $x_i$  to be a simple variable notation which is  $z_i$ , so this form becomes a bit simpler and now we know of course if we can model probability of  $y_i$  equal to 1, we can also talk about probability of  $y_i$  equal to 0, the complement of the event that there is a success.

So of course there will be a failure, so here we are going to talk about the negative outcome or the failure and it is important because you will see later on when we are going to talk about interpretation of logistic models. We have to have very clear idea what probability we are modeling because you know odds ratio is basically a ratio of probability of two events in a success and a failure.

So it is important to know how we know we can also calculate the probability of a negative outcome or a failure. So if I say that  $y_i$  is equal to 1 actually indicates a success, then  $y_i$  equals to 0 will indicate failures, say and that can be computed very easily as we have this number for  $\pi_i$  from the logistic distribution. So  $1 - \pi_i$  will give me the probability of the negative outcome or probability that  $y_i$  will take 0 value.

(Refer Slide Time: 13:48)



**The Logistic function**

- ▲ However, estimation of  $P_i$  becomes complicated because  $P_i$  is nonlinear not only in  $X$  but also in the parameters  $\beta$ .
- ▲ In order to avoid such complication, we use a concept called **odds ratio (OR)**.
- ▲ The OR is defined as the ratio of probability of success over probability of failure

$$\frac{P_i}{1 - P_i} = \frac{1 + e^{z_i}}{1 + e^{-z_i}} = e^{z_i}$$

- ▲ If we take natural logarithm of the OR, we get what is called **logit**

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = \ln(e^{z_i}) = z_i = \beta_1 + \beta_2 x_i$$

here  $L_i$  is the log of the OR which is linear both in  $X$  &  $\beta$ 's.

- ▲ Note that as  $P_i$  goes from 0 to 1, the  $L_i$  goes from  $-\infty$  to  $+\infty$ . This implies that only probability is bounded between 0 & 1, but not the *logit*.

So, now with this let me move on and then the estimation, we are going to talk about in this particular slide. We will start talking about the estimation but first, let me introduce the odds ratio in this context. So note that the estimation of pi becomes very complicated because pi is non-linear not only in the explanatory variable but also in the parameters beta.

So we have to somehow find the solution to this highly non-linear model and that is where we are going to make use of this odds ratio concept that we have studied in the first part of the course. So here we are going to introduce the odds ratio as ratio of two probabilities. So odds ratio is defined as the ratio of probability of success over the probability of failure. So it is given by pi divided by 1 minus pi.

So I am showing here that if you do some small algebra you get a very simple expression for that and that is e to the power zi. So, but still this is exponent, there is no linear function nothing that we can deal with in regression analysis. So we take natural logarithm of the odds ratio and we get what is called a logit and this logit is given by Life insurance.

So, if you take the natural logarithm both sides, you finally get that logit is equal to zi which is equal to beta 1 plus beta 2 times xi. So, now you see these probabilities are not linear functions of explanatory variable or the parameters in the regression model but the logic function that we have defined just now it is a linear function of the explanatory variable.

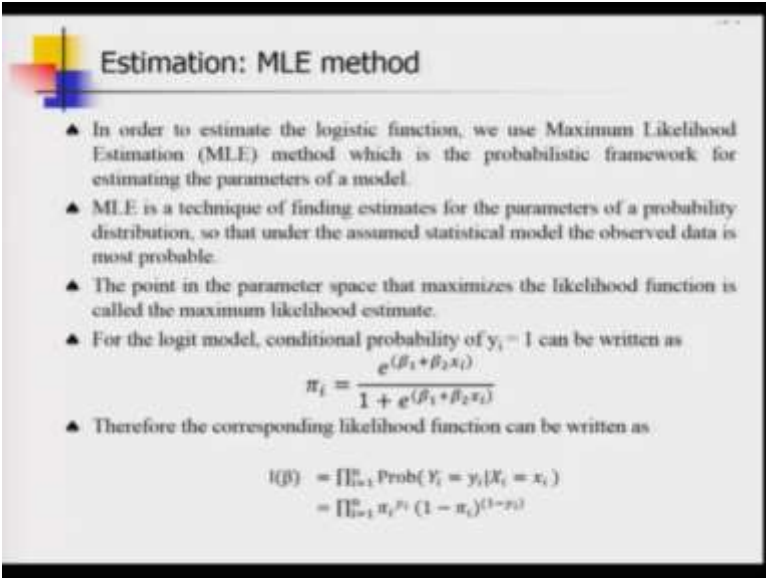


So this logit function is linear in the explanatory variable and in the parameters also. Note that as the probability of success  $\pi_i$  goes from 0 to 1, the value of logit function goes from minus infinity to plus infinity. So, this implies that only the probability is bounded between 0 and 1 but not the logit function.

So we have an equation which is linear in explanatory variable and linear in parameter. So what is stopping us to adopt OLS here? See this is a very interesting case, here you have  $y$  which is not observed because what is the dependent variable here? It is no longer 0 and 1, the outcomes of this qualitative variable that you can actually observe. What is the dependent variable? This is the logit and logit involves the probability of success.

Now we do not know, what is the probability of success? So there is no data on probability of success for different individuals in the sample. So you have a dependent variable which is unobserved. So you cannot apply the standard OLS technique and we have to apply what is known as Maximum likelihood estimation method.

(Refer Slide Time: 17:11)



**Estimation: MLE method**

- ▲ In order to estimate the logistic function, we use Maximum Likelihood Estimation (MLE) method which is the probabilistic framework for estimating the parameters of a model.
- ▲ MLE is a technique of finding estimates for the parameters of a probability distribution, so that under the assumed statistical model the observed data is most probable.
- ▲ The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.
- ▲ For the logit model, conditional probability of  $y_i = 1$  can be written as
$$\pi_i = \frac{e^{(\beta_1 + \beta_2 x_i)}}{1 + e^{(\beta_1 + \beta_2 x_i)}}$$
- ▲ Therefore the corresponding likelihood function can be written as
$$l(\beta) = \prod_{i=1}^n \text{Prob}(Y_i = y_i | X_i = x_i) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)}$$

So in this slide I am going to give you a brief introduction to Maximum likelihood estimation method. So, first of all, what is it? So, it is a probabilistic framework for estimating the parameters of a model. So note that when we talked about estimation in statistical inference in Part 1 of the course, I have there only spoken about this particular method and I said that I will

come back later in the course with an example or a case where you know Maximum likelihood estimation method is to be applied and today is the day I am going to talk about that method. So note this important fact that in statistics and econometrics, when we say estimation, we do not mean that we have to always OLS.

OLS is not applicable to all possible cases that we see around us. so here we are going to talk about this new method that is an alternative estimation procedure and if you remember I have also spoken that once you decide to apply Maximum likelihood estimation procedure, you have to have an assumption regarding the distributional form of the stochastic random disturbance term, which you do not have to assume in the case of OLS. So I hope that you will remember this distinction between MLE and OLS.

Now let us dig deeper and see what MLE actually does for us. So the second bullet point and the third bullet point in this slide is going to discuss or show what MLE is doing for us. So MLE is a statistical technique of finding estimates for the unknown population parameter values of a probability distribution, so that under the assumed statistical model, the observed data is most probable. So what do we mean by that?

So, we mean that when you draw a sample from a population, that population always will have an underlying probability distribution for that particular variable for which you are conducting statistical data analysis and you do not know a priori which population or probability distribution is there from where I am actually observing some realizations of values.

So of course there you need to make an assumption be it normal, be it logistic, be it gamma whatever distribution but once you make that assumption that in most possible cases this variable in population follows this particular distribution, then basically the task is to find the parameter values of that probability distribution, so that you can compute the probabilities and all.

So, Maximum likelihood estimation procedure actually helps you to find out the proxies or the estimates for those unknown and unobserved population parameter values such that the probability that the sample has come from probability distribution with these parameter values is the highest. So basically what I am trying to say here, let me explain it in different words. So suppose you know that this sample that you have at hand has come from a normal probability

distribution but you do not know the values of the mean and the variance of this population probability distribution.

So you have to get some proxies or estimates for  $\mu$  and  $\sigma^2$  the probability distribution which is embedded in the population. Now there could be many values of  $\mu$  and  $\sigma^2$  you do not know which one is the true value of the population parameter. So then you try with a set of  $\mu$  and  $\sigma^2$  and then if you say that okay I have this value  $\mu_1$   $\sigma_1^2$  and then I have another set  $\mu_2$   $\sigma_2^2$ , then I have another set  $\mu_3$   $\sigma_3^2$ . These are all alternative values of  $\mu$   $\sigma^2$  combination that can actually take place.

Now once you assume a particular set of values for  $\mu$  and  $\sigma^2$ , then you can compute the probability that your sample has come from normal distribution with this particular value of mean and  $\sigma^2$  and you are then going to compare all these probabilities that you can theoretically calculate for all these possible values of  $\mu$  and  $\sigma^2$  and MLE actually a very special type of estimation process which will find out that particular value of  $\mu$  and  $\sigma^2$  which will actually maximize this probability that you are calculating by plugging these proxy values of  $\mu$  and  $\sigma^2$ .

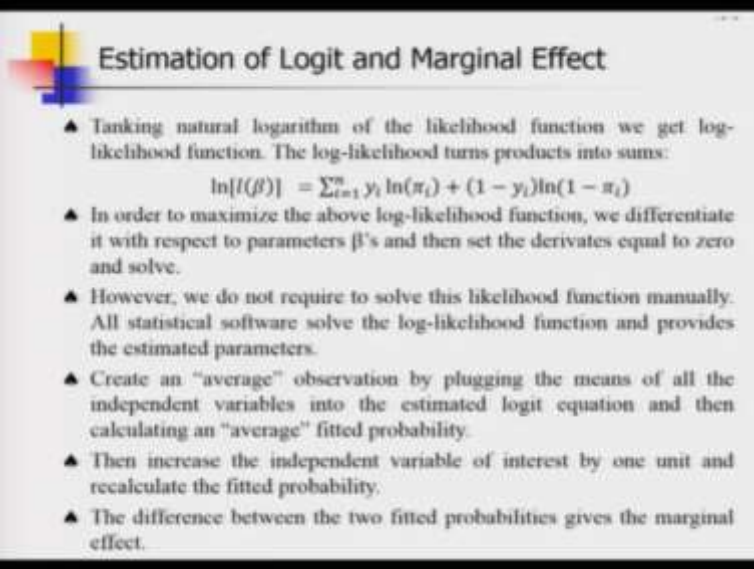
So Maximum likelihood estimation method actually maximizes the likelihood which is basically nothing but probability and it maximizes this objective function in parameter space. So the solution point actually is called the Maximum likelihood estimate. So now let us come back to logit model again and then let us see how MLE is going to be applied to logit model. So for the logit model, we write the conditional probability of  $y_i$  equals to 1.

So why conditional probability? Because it is conditional upon the values of  $x_s$  the explanatory variables. And the conditional probability can be written as  $\pi_i$  and then you see the expression is given and then we can write the likelihood function as  $L$  of  $\beta$  and you see I am multiplying all these conditional probabilities. Why am I multiplying? Because I am assuming that these probabilities are all independent of each other.

So as they are independent, they could be multiplied together. And then what I am doing? You see I get a simpler expression in terms of  $\pi_i$ , note that I am skipping some steps in this theoretical

calculation because ultimately this is not a theory class and we are not going to pay much attention to proofs. So you know I am going to skip couple of steps, so that I save time.

(Refer Slide Time: 24:21)



**Estimation of Logit and Marginal Effect**

- ▲ Taking natural logarithm of the likelihood function we get log-likelihood function. The log-likelihood turns products into sums:
$$\ln[l(\beta)] = \sum_{i=1}^n y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)$$
- ▲ In order to maximize the above log-likelihood function, we differentiate it with respect to parameters  $\beta$ 's and then set the derivatives equal to zero and solve.
- ▲ However, we do not require to solve this likelihood function manually. All statistical software solve the log-likelihood function and provides the estimated parameters.
- ▲ Create an "average" observation by plugging the means of all the independent variables into the estimated logit equation and then calculating an "average" fitted probability.
- ▲ Then increase the independent variable of interest by one unit and recalculate the fitted probability.
- ▲ The difference between the two fitted probabilities gives the marginal effect.

Now you take natural logarithm of the likelihood function and then we get the log likelihood function. So the log likelihood actually turns the products into sums because it is easy to deal with sums. So the expression is given, its quite complicated and then you actually maximize the log likelihood expression that I have shown there in the slide with respect to the unknown parameter values betas like we do in the case of OLS.

and I think you as of now, this is good enough of our discussion on Maximum likelihood estimation because it is very difficult to do this kind of exercise on a piece of paper for logistic regression but fortunately softwares are there, so if you ask a software it will happily do it for you within a minute.

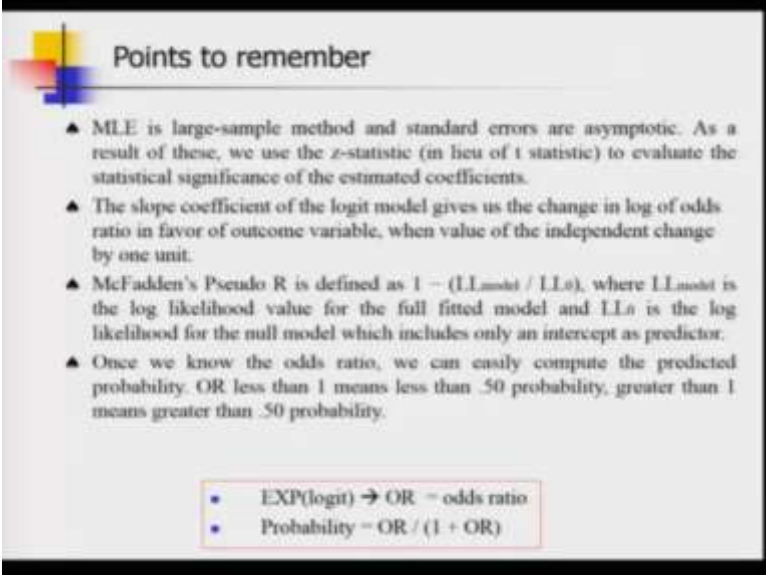
So let us not waste time on this Maximum likelihood estimation aspect of logistic regression model. Let us move ahead. So once some estimation is been done, then we are interested in the marginal effect. So you remember that in the case of OLS also, we actually attached this marginal effect tag with this regression coefficients because the slope coefficient in the OLS regression context has this marginal effect interpretation.

But in the context of logistic regression, marginal effect has a very different kind of interpretation. So let us have a look at this thoroughly and the complication arises because now you do not have a simple  $y$ , so you have a logit as the dependent variable. So if there is a change in explanatory variable, it is going to impact the logit and not the original  $y$  variable that you have observed.

So what we have to do? We have to create an average observation first by plugging the means of all the independent variables into the estimated logit equation. So here we are assuming that the logit equation that we have estimated involves only continuous explanatory variables. If there is dummy then you have to plug either 0 or 1 because you cannot take mean of that and if you do so then, you get a measure of average fitted probability.

And then you can increase the independent variable of interest by holding others fixed just by one unit and then recalculate the fitted probability again and the difference between these two fitted probabilities gives us the marginal effect of that explanatory variable, provided the other explanatory variables are all kept constant.

(Refer Slide Time: 27:23)



**Points to remember**

- ▲ MLE is large-sample method and standard errors are asymptotic. As a result of these, we use the  $z$ -statistic (in lieu of  $t$  statistic) to evaluate the statistical significance of the estimated coefficients.
- ▲ The slope coefficient of the logit model gives us the change in log of odds ratio in favor of outcome variable, when value of the independent change by one unit.
- ▲ McFadden's Pseudo  $R$  is defined as  $1 - (LL_{model} / LL_0)$ , where  $LL_{model}$  is the log likelihood value for the full fitted model and  $LL_0$  is the log likelihood for the null model which includes only an intercept as predictor.
- ▲ Once we know the odds ratio, we can easily compute the predicted probability. OR less than 1 means less than .50 probability, greater than 1 means greater than .50 probability.

- $EXP(\text{logit}) \rightarrow \text{OR} = \text{odds ratio}$
- $\text{Probability} = \text{OR} / (1 + \text{OR})$

Now, before we end our discussion on the theory of logistic regression, we have to keep certain points in mind and these are all important points. So here I am going to talk about the first important point which is a major departure from the OLS that we have been doing so far. so here

for hypothesis testing for individual regression coefficients, we are going to look at the z scores and not the t scores.

Why? Because t actually is a small sample case but as MLE assumes that we are dealing with large sample and asymptotic properties are holding well for MLE then we have no problem, we can straight away make use of the z statistics. So that is why in the software outcome you will see that softwares are always reporting the z statistic and you have a look at that z score and the corresponding p value and you can take a decision whether this particular coefficient is significant or not and remember that I have spoken about this magic value of 1.96.

That holds here also for the case of z. So if you see the z value is above that number, then you can assume that that particular regression coefficient is statistically significant. So the second point that we must remember is that the slope coefficient of the logit model gives us the change in log of odds ratio in favor of the outcome variable and this is the case when the value of the independent variable change just by 1 unit.

So the next point we are going to discuss is the goodness of fit case. So if you remember in the context of OLS regression, we have introduced two goodness of fit measures and they are r square and adjusted r squares. So here in this context you may be interested to know how do I know that I have good model fit. Now this is a fairly complicated model, so if you are thinking of having some simple measure, it is not difficult to propose one and that is why here you will not get r square or adjusted r square like we were used to get in the case of OLS.

So here luckily some econometricians have come up with some measure and they say that this is a pseudo measure of r square, the coefficient of determination and now we are going to talk about one such measure, there are many pseudo measures available for r square and here in this lecture I am going to talk about only one of them and that is proposed by a Nobel laureate econometrician Daniel McFadden.

So there is a typo here, so there has to be a square term, please make a note of that. So McFadden's pseudo r square is defined as  $1 - \frac{\text{LL model}}{\text{LL null}}$  and the ratio is basically of two log likelihood numbers and what are these two log likelihood numbers? LL model is the log

likelihood value for the fully fitted model and LL0 is the log likelihood for the model under null hypothesis, so that means that no explanatory variables are affecting or determining  $y$ .

So it is basically an intercept only model. So basically you have to run the logit model twice, once with all possible explanatory variables that you have in your regression equation and you get one log likelihood number and that is your LL model and then you run another logit model but this is an intercept-only model and you will definitely get another log likelihood value for it.

So you take a ratio of that and deduct this ratio number from 1 and you are going to get McFadden  $r$  square. So once we have the estimated coefficients, we can always find out the odds ratio from the logit function and from there we can easily compute the predicted probability. So-- but odds ratio number itself gives us a signal about the probability of the positive outcome, so odds ratio less than 1 means that there is less than 50 percent probability and if odds ratio is greater than 1 then that indicates greater than 50 percent probability.

So here by odds ratio I am always meaning that you know we are talking about ratio of probabilities, probability of success divided by probability of failure. Now at the bottom of the slide, I have reported two important formula that you must remember and they are given inside this red box. So they are basically summarizing what we have discussed in bullet point 4 only.

Let us stop discussing the theory of logistic regression right here only because the extensions are quite difficult to appreciate and I have to introduce many-many new concepts, if I want to explore other things like hypothesis testing, confidence interval etc. So at this point I have decided to stop talking about the theory of logistic regression.

I hope that although the Maximum likelihood estimation part was a bit difficult one to appreciate but you will at least get a fair idea why we require logistic regression models and how to execute that model, especially if you are using a software then what is actually happening behind the curtain and once the software throws a set of numbers at you, how to interpret them.

What use you can make of them, so that if you understand from this lecture that is good enough and I believe that from my perspective objective is made that I just wanted to give you an idea about a model which is practical but it is somewhat different from what we usually do in this typical linear regression model setup.

So now it is time to talk about case studies because this logistic regression has been applied in empirical research a lot and in different fields. So I am going to talk very briefly maybe two three minutes on two case studies, so that you get a feel that how in India this logistic regression model has been applied by researchers to you deal with different problems.

(Refer Slide Time: 35:44)

**Case study 1: Factors affecting birth weight of newborn babies in rural India**

- ▲ Objective: "To identify the factors affecting the birth weight of newborn"
- ▲ Dependent variable: Indicator for Low Birth Weight (LBW) which is takes value 1 if the newborn baby has LBW and 0 otherwise.
- ▲ Independent variables:
  - Maternal education [5 categories]
  - Exposer to passive smoking [3 categories]
  - Age at first pregnancy [3 categories]
  - Birth interval [2 categories]
  - Previous LWB babies [continuous]
  - Weight gain during pregnancy [4 categories]
  - Maternal weight at last week of gestation [2 categories]
  - High risk pregnancy [3 categories]
  - Antenatal registration [4 categories]

Source: Metgud, C. S., Naik, V. A., & Mallapur, M. D. (2012). Factors affecting birth weight of a newborn- a community based study in rural Karnataka, India. *PloS one*, 7(7), e40040.

So the first case study is going to come from the field health economics, you can also say that medical science literature because there are so many intersections and overlaps between these two fields. So here I am going to talk about a paper which got published in an international journal. So, now let us talk about the paper that I mentioned, so it is a paper published by Metgud Naik and Mallapur in the year 2012 and it got published in an international journal PloS one.

So in this paper, the authors have identified the factors affecting birth weight of a newborn child and this case study has come from rural Karnataka. So the authors start with this objective in the paper that they want to identify the factors affecting the birth weight of a newborn. So they have collected data on the weights of newborn babies in one region of rural Karnataka and they have got a very large sample and then they actually defined the dummy variable and they used code 1 for babies where they observed that the birth weight of the newborn baby is less than the average or what should be called weight of a good healthy newborn baby.



So if it is less than that then they have called that the dummy variable LBW, no LBW will take value 1. So here the dependent variable is dummy variable, so this indicator variable LBW will take value 1, if the newborn baby has a birth weight which is less than or lower than the standard weight that doctors consider to be good weight of a healthy newborn and otherwise this dummy variable takes 0 value.

Now the authors have introduced a set of independent variables to model this LBW variable and here we have shown you the list of those variables, note that only one variable here is continuous that is the previous number of low birth weight babies other than that they are all qualitative variables. So of course there are different categories and for different categories dummy variables were used to take care of these factors.

So as we see there are many-many dummy variables in the regression model and that is why it is not possible for us to talk about the effects of all these dummy variables or to show you the regression result for all those dummy variables, so we have decided that we will talk about two important qualitative factors which have affected this particular variable LBW pretty much. So in the next slide, I am going to show you results regarding those two qualitative or attribute variables only.

(Refer Slide Time: 39:11)

**Case study 1 (contd.)**

◆ Among this list independent variables, here we present odds ratio of two most important qualitative variables

Variable name	Univariate odds ratio	Multivariate odds ratio
<b>Maternal education</b>		
Illiterate	2.4 (1.4 – 4.1)*	3.2 (1.6 – 10.3)***
Primary education	1.5 (0.9 – 2.4)	2.9 (1.1 – 8.0)***
Secondary education	1.1 (0.7 – 1.7)	1.9 (0.7 – 4.9)
Pre university	Ref. category	Ref. category
<b>Exposure to passive smoking</b>		
No	Ref. category	Ref. category
Husband	2.0 (1.3 – 3.2)**	2.3 (1.1 – 4.9)***
In-laws & others	1.1 (0.7 – 1.8)	1.2 (0.6 – 2.8)

Note: 95% confidence interval are given in parenthesis

Now here I am going to talk about the odds ratio numbers. So here you see that first column of the table shows the variable names, so if maternal education is one qualitative variable you see that here the authors have used different dummy variables to take care of that. So pre-university has been declared the reference category or the base category.

So there were three dummy variables after excluding that category and they are illiterate, primary education and the secondary education. So secondary education variable means that the mother has attended secondary education and then the second qualitative or attribute variable which we are interested in is exposure to passive smoking and this has three categories or labels associated with it and if the mother was not exposed to passive smoking during the pregnancy period or the pre-pregnancy period then we call that as the reference category or the base category, so that no category will be excluded, there will be no dummy for it but there are two dummy variables namely husband and in-laws and others.

So what do they mean? So husband dummy means that this pregnant woman's husband, he used to smoke and she was exposed to passive smoking from the husband and similar interpretation can be given to the in-laws and others dummy variable as well. Now you see we have the second column and third column, there we are seeing odds ratio and the confidence intervals, we are not going to talk about the confidence intervals because we have not studied it in this logistic regression context but let us talk about what do we mean by univariate odds ratio and the multivariate odds ratio.

So in the paper the authors have calculated univariate odds ratio by the following approach. Suppose, we are interested about one particular variable, say husband dummy. So then they have run a model with that husband dummy and the intercept only to get the odds ratio for the husbands and if suppose they are interested in the illiterate dummy to see the impact of the illiterate dummy, then they have run the logistic regression with intercept and this illiterate dummy in the regression equation to note down the odds ratio of that particular dummy variable illiterate.

So they have then run series of logistic regressions each time including just one explanatory variable and intercept of course was there and that is the way they have computed this univariate odds ratio. Now what is the multivariate odds ratio? So basically now this is just one single piece

of regression where all variables were thrown to the software together and there is one regression output from which you can now look at the effect of a particular individual explanatory variable while all others are sitting there. So now let us have a look at the table again.

So you see one interesting thing here, so if we compare the univariate odds ratio and the multi weight odds ratio for all the variables you see that the multivariate order ratio are higher compared to the univariate odds ratios. Now let us concentrate on one particular number and then let us see know whether we can provide a good interpretation for that number or not.

So suppose we concentrate on the husband case, okay so if we are interested in the case of husband, say let us you know consider the final model which is the multivariate odds ratio column. So from there we see that the odds ratio is 2.3 and it is highly significant. So what do we mean by this number 2.3? So remember that when we talked about odds ratio I said that think of ratio of probabilities, so probability of  $y_i$  equals to 1 divided by probability of  $y_i$  equals to 0.

So here  $y$  is what? That is your LBW or LWB that low birth weight babies delivery, so that is basically the event that you are modeling. So basically here the odds ratio is talking about probability that the newborn baby's low weight compared to the probability of the newborn baby above the standard weight.

So you see that if a pregnant woman's husband is a smoker, then it is very risky because the probability that the women will deliver baby with low birth weight that is 2.3 times higher than women whose husband does not smoke and she is not exposed to passive smoking from husband. Now let us look at the role of education because education plays a very big role in socio-demographic research.

So let us have a look at the what education is telling us here. So here you see that from illiterate to primary education that means that the mother is little bit educated, then primary to education to second education means that mother is more educated.

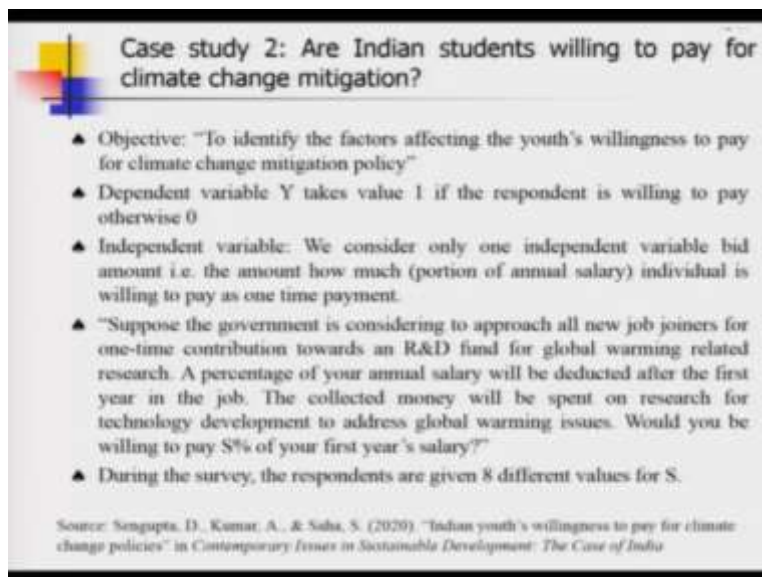
So here you note down the odds ratio for this consecutive dummy variables. So as one can expect that more education is beneficial for not only the mother but also the baby because then mother knows about best practices to follow and probably take care of the baby inside during the

pregnancy period, so as the education level of the mother goes up, we can expect that the risk of delivering a low weight baby will fall.

So that is what is reflected in the declining numbers of the odds ratio as well. So you see here for an illiterate women the multi-weight odds ratio number is 3.2. So that means that the chance is 3.2 times higher that particular illiterate women will give birth to a low weight baby but note that if I now consider the multiweight odds issue 1.9 and that is for secondary education, so that means that for a mother who has attended secondary education, this risk has come down drastically.

So here there is 1.9 times chance that this mother will give birth to a low weight baby. So what we learnt from this discussion on education? That proper education of women it reduces the risk of giving birth to a low weight baby.

(Refer Slide Time: 48:01)



**Case study 2: Are Indian students willing to pay for climate change mitigation?**

- ▲ Objective: "To identify the factors affecting the youth's willingness to pay for climate change mitigation policy"
- ▲ Dependent variable Y takes value 1 if the respondent is willing to pay otherwise 0
- ▲ Independent variable: We consider only one independent variable bid amount i.e. the amount how much (portion of annual salary) individual is willing to pay as one time payment.
- ▲ "Suppose the government is considering to approach all new job joiners for one-time contribution towards an R&D fund for global warming related research. A percentage of your annual salary will be deducted after the first year in the job. The collected money will be spent on research for technology development to address global warming issues. Would you be willing to pay S% of your first year's salary?"
- ▲ During the survey, the respondents are given 8 different values for S.

Source: Sengupta, D., Kumar, A., & Saha, S. (2020). "Indian youth's willingness to pay for climate change policies" in *Contemporary Issues in Sustainable Development: The Case of India*

Now we move on to the second case study and this one I have picked from the field of environmental and climate change economics because nowadays there is lot of talk regarding environmental pollution and global warming and this that this is a study that was published in a book which got published in 2020 and this is a chapter from that book written by Sengupta Kumar and Saha and the title of the chapter reads as Indian youth's willingness to pay for climate change policies.

Now willingness to pay is a very popular and well known concept in the field of public economics and environmental economics but if you are not coming from that background let me educate you with couple of sentences. So basically suppose we are talking about creation of a public good which is not to be consumed by one particular individual, so this good is basically a common good.

So suppose if we go for afforestation on a vast piece of barren land, that land is basically with the government so that land is not owned by anyone. So actually people will not derive any private benefit out of that afforestation project but that does not mean that society as a whole will not gain some positive benefits from that project.

So, if you now go to the community and conduct a survey and ask people that okay there is something good happening and are you willing to pay something for this positive activity say afforestation? Then the person may say yes may say no and there are many-many factors why a particular person is saying yes or no. So you in a nutshell without getting into the theoretical complications, this is willingness to pay all about.

So now we are going to visit what the authors have done in this particular chapter. So the objective was that they wanted to identify the factors that affect the youth's willingness to pay for climate change mitigation policy. So of course government of India if it wants to take drastic measures to protect us from the future global warming scenarios. So we have to take mitigation policies but these mitigation policies are all costly, nothing comes for free right.

So now the issue is that whether the society is ready to sacrifice some bit of their current resources to future to protect their future generations that is the question. So here the paper or the chapter actually wanted to address this particular question by conducting a survey amongst the students of various engineering colleges in the Kanpur region and the dependent variable here  $y$  takes a value 1, if the respondent is willing to pay and 0 otherwise. So now what are the independent variables?

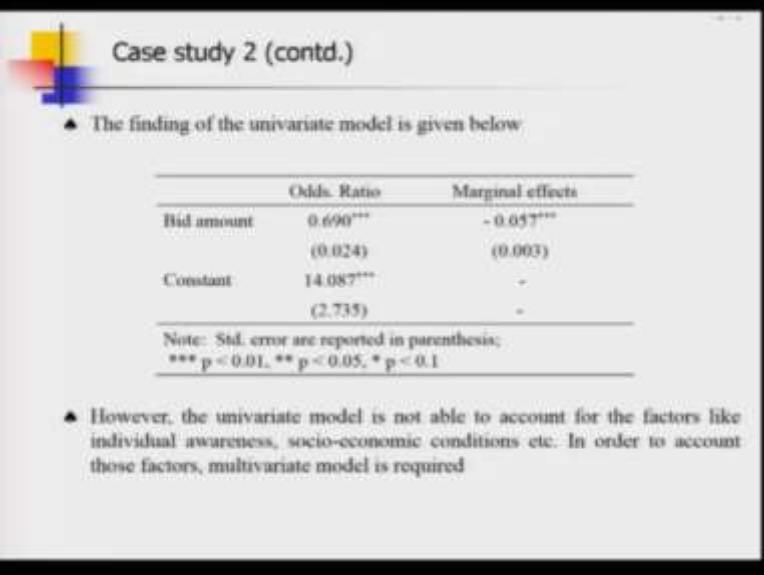
So that chapter has actually used three four explanatory variables but due to shortage of time I just wanted to talk about one particular variable which is the bid the the amount of money that

the respondent was asked to sacrifice in order to help the governments in targets to protect global protect us from global warming and roll out climate change mitigation policies.

So here we are talking about only one explanatory variable and now let us look at the question that was asked. So here I have shown you this question here so you see that the respondent was asked to sacrifice or part with S percent of first year's salary and there were different eight different values of S or this percentage of salary sacrifice that was asked to respondent.

So one respondent got to see one particular value of S only and based on that he or she said yes or no. So suppose I have chosen a value one for one particular respondent, so basically the respondent was asked whether he or she is ready to part with 1 percent of the first year salary and the respondent has this option to say yes or no. So this is basically the model setup.

(Refer Slide Time: 52:47)



Case study 2 (contd.)

- ▲ The finding of the univariate model is given below

	Odds Ratio	Marginal effects
Bid amount	0.690*** (0.024)	-0.057*** (0.003)
Constant	14.087*** (2.735)	- -

Note: Std. error are reported in parenthesis;  
\*\*\* p < 0.01, \*\* p < 0.05, \* p < 0.1

- ▲ However, the univariate model is not able to account for the factors like individual awareness, socio-economic conditions etc. In order to account those factors, multivariate model is required

So here I am showing you the results but note that here as this is a continuous variable you can measure the marginal effects and I have explained already how to calculate marginal effects odds ratio is also given so here I end this discussion because we are running out of time. So we are done with our discussion on discrete choice models and in the next week, I am going to come back with some new topics so see you then. Thank you. Bye.