Handling Large Scale Unit Level Data Using STATA

Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture 34

Binary Response Models - III

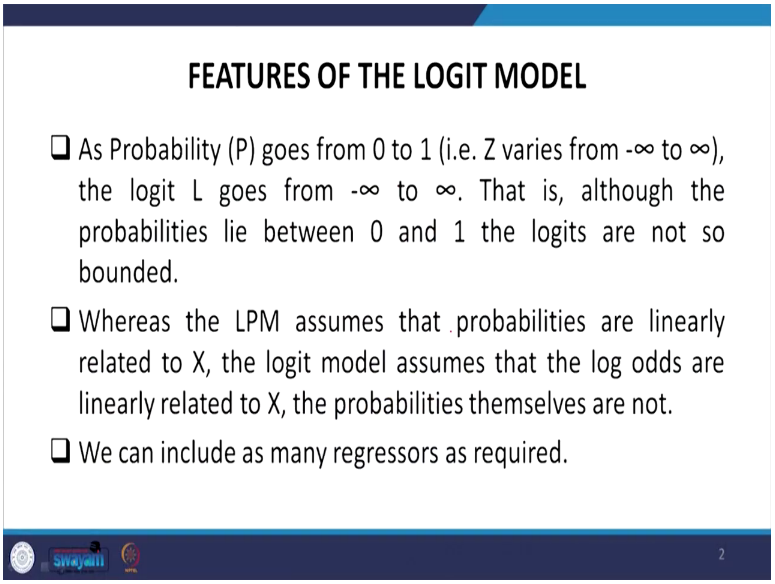
Welcome friends to this NPTEL MOOC schedule or course on Handling Large Scale Data Using STATA. We are at the particular lecture of understanding binary response models. In the last lecture we already started discussing about Logit model and the theory. The basic premises behind the Logit as against the linear probability model where OLS has been contradicted by the Logit. We are going to validate further also in this particular lecture.

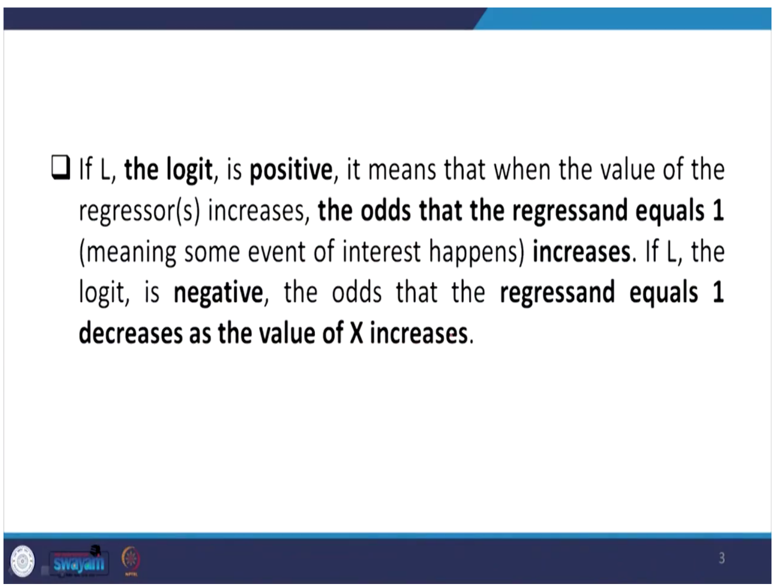
So, today we will emphasize further some other features of Logit, particularly from the last class if you remember correctly that the equation we are emphasizing on the log of odds ratio. That is particularly defining a linear equation in terms of the expected coefficients. That we mentioned.

We are adding some attributes to the Logit model. Even I said earlier, when the probability in the diagram I told you already, it is a sigmoid curve Logit formation follows a sigmoid graph where the probability's value varies from 0 to 1 with the Z value varies from minus infinity to infinity.

So, the Logit that is L goes from minus infinity to infinity and though the probabilities lie between 0 to 1 the logits are not so bounded. So, value in the STATA also we are not going to get any bounded values. We should not be just confused by their coefficients, that's why it is so, why it is not so? We will get to know it for sure.

(Refer Slide Time: 02:42)

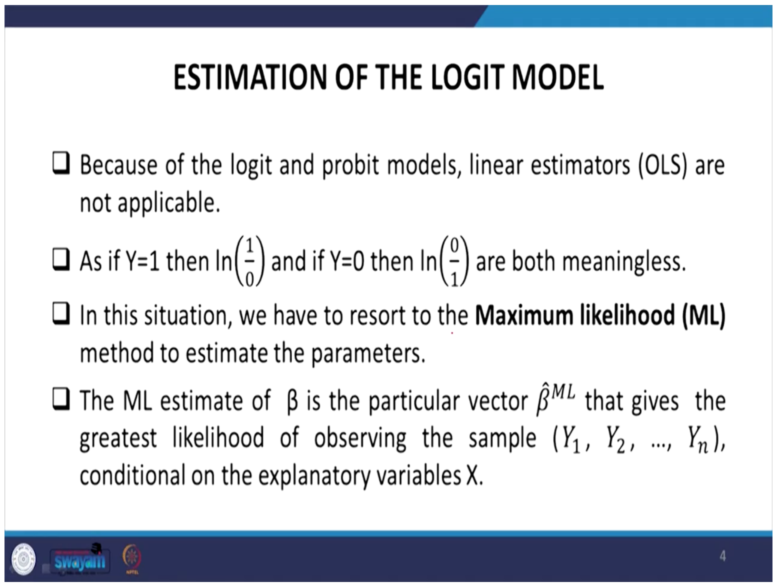




Whereas the LPM assumes that the probabilities are linearly related to X. The Logit model assumes that the log of odds are linearly related to X. The probabilities themselves are not necessarily linear. We can include as many regressors when so required for the model. So, the L that stands for Logit is positive, means that when the value of the regressor increases the odds of that regressand equals 1, meaning some event of interest happens, also increases.

If L, the Logit is negative the odd of that regressand equals 1 decreases as the value of X increases. So, we are going to also present that odds ratio in terms of comparing with the unit value. If it is exceeding unit value then it indicates a positive signal. That means the regressor impact the likelihood of the dependent variable positively. If the odds ratio is less than 1 then it is negatively impacting. We are going to emphasize this with help of data.

(Refer Slide Time: 04:13)



Because the Logit and Probit models and its linear estimation that is OLS are not applicable, so the estimation of the Logit model is different. And since the model itself is not linear so OLS is not applicable in this case. If Y is equal to 1 then the probability with 1 or 0 as the odd ratio. And if it is 0 then the odd ratio is 0 by 1 are both meaningless.

In this situation we have to resort to the maximum likelihood. It is not just the probability with 1 and 0; rather the closeness to 1 or closeness to 0 that is identified through the maximum likelihood estimator is more important than just the probabilities. So, the ML estimate of the beta is a particular vector that is beta hat ML that gives the greatest likelihood of observing the sample of Y_1, Y_2 till Y_n conditional upon the explanatory variable X.

When we say the unit change within the probability limit of 0,1 is very important and the maximum likelihood estimator identifies that so the coefficient itself is not just enough. The absolute coefficient which is scattering to probability of success or failure or odd of success or not success is not enough.

(Refer Slide Time: 06:07)

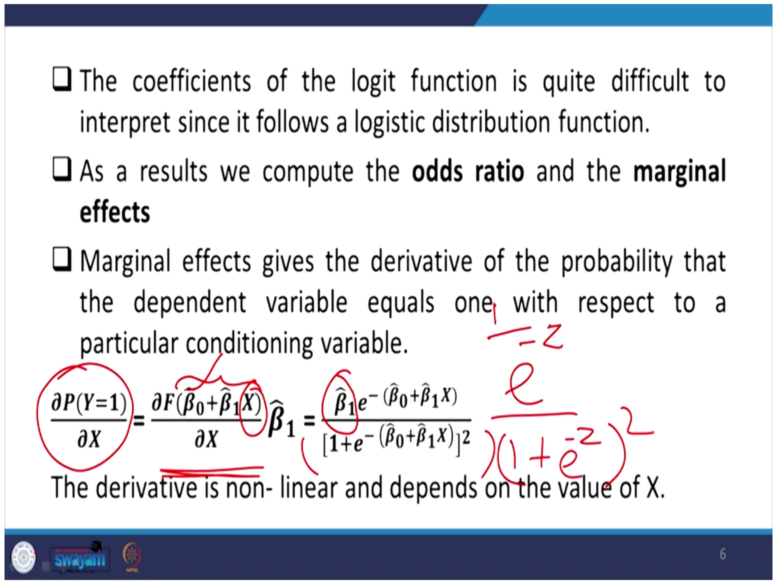
MARGINAL EFFECTS IN LOGIT MODEL

- In linear regression model, the slope coefficient measures the change in the average value of the regressand for a unit change in the value of a regressor.
- The actual coefficients in a logit or probit analysis are limited in their immediate interpretability.
- The signs are meaningful, but the magnitudes may not be, particularly when the variables are in different metrics.
- We can not interpret the coefficients directly in terms of units of change in Y for a unit change in X, as in regression analysis.

- The coefficients of the logit function is quite difficult to interpret since it follows a logistic distribution function.
- As a results we compute the **odds ratio** and the **marginal effects**
- Marginal effects gives the derivative of the probability that the dependent variable equals one with respect to a particular conditioning variable.

$$\frac{\partial P(Y=1)}{\partial X} = \frac{\partial F(\hat{\beta}_0+\hat{\beta}_1 X)}{\partial X}\hat{\beta}_1 = \frac{\hat{\beta}_1 e^{-(\hat{\beta}_0+\hat{\beta}_1 X)}}{[1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X)}]^2}$$

The derivative is non- linear and depends on the value of X.



Rather in that case the unit change value or the maximum likelihood estimator of its unit change expressed through marginal effects is very important. Since the dependent variable is with 1 and 0 means binary so the unit change the very fractional change and its interpretation is very relevant. So, that is the reason why we are sticking to marginal effects.

In linear regression model the slope coefficient measures the change or the average value of the regressand for the unit change in the value of the regressor. The actual coefficients in the Logit or Probit analysis are limited because of their immediate interpretability. The signs are meaningful but the magnitude may not be. The magnitude of the coefficient is not meaningful particularly when the variables are in different metrics.

We cannot interpret the coefficients directly in terms of units of change in Y with the unit change in X in the regression analysis. Rather the coefficient of the Logit function is quite difficult to interpret since it follows a Logistic distribution function. It is not a continuous series. There are fractional change in the probabilistic structure with the change in the Z values from minus infinity to infinity. So, as a result of this we compute the odds ratio and also the marginal effects.

Marginal effects give the derivative. Basically when we say marginal that means we have to take the derivative of the dependent with respect to the derivative of the independent. So, in the independent we have X as the set of, or vector of independent variables, the regressors. So, if you take the partial derivative because we have so many independent variables, you take partial derivative with respect to X here then in that case it is important to find out, so it is important to locate how partial it is and with the unit change in this.

So, let me just fix the pen here. Whether it is coming? It was there earlier. Let me just check. I hope it is not been set. It is probably here alright it is the pen mode. So, touch is active, alright. So, I will operate through this. Here now I am discussing about the partial derivative of the dependent variable with respect to the independent variable.

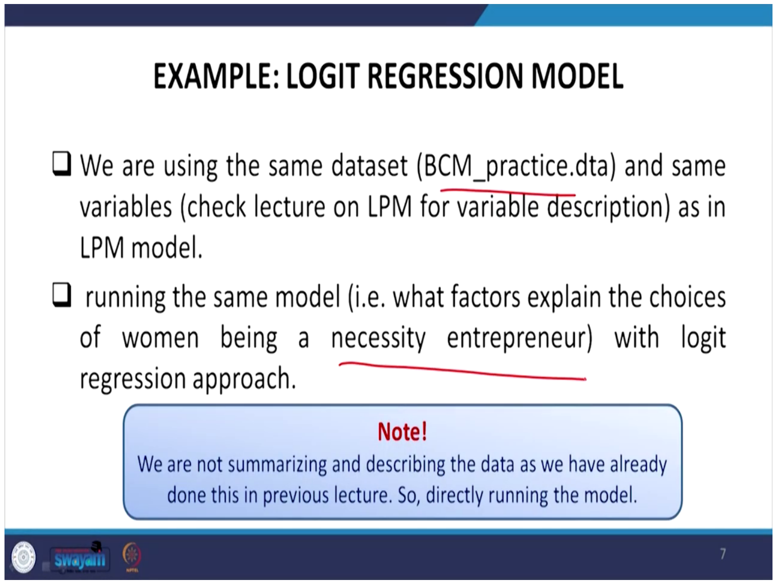
When we take the first derivative it will be certainly be of the dy/dx in the right hand side. You look at this carefully. dy/dx, dy is nothing but the cumulative distribution function, logistic distribution function with this. And when we take it, since X is here and this is a constant alright. So, the derivative will be, the others are constant so beta 1 hat is left. That will be simply multiplied.

After that, what is that function of that logistic? We have already mentioned that beta hat plus beta 1 X is nothing but in the logistic function This is separated and we are left with. So, the function is e to the power minus Z. So, 1 upon E to the power minus Z.

So, what we will do? If you take the derivative it will be like here. If I just take the derivative, it will be, in the denominator it will be if it is in terms of ratio the denominator will be simply square of it, in the first derivative will be square of it. Then basically e, that is in terms of square is there in the denominator and in the numerator it is simply e to the power minus Z, e to the power minus Z and 1 plus this term, this is of square. So, this is what we have derived.

The derivative is in fact nonlinear and depends on the value of X. So, what is the meaning of it? That means the marginal effect or the impact of this we wanted to find out is whatever the marginal effect of Y observed through the X. So, Z here we already defined in the last class that it is dependent upon X. Coming to the example, the real life example of testing Logit function, the Logistic estimation; there are slight difference between Logistic and Logit because of the ratio or the coefficient we are going to estimate. That we are mentioning in a short while.

(Refer Slide Time: 11:46)

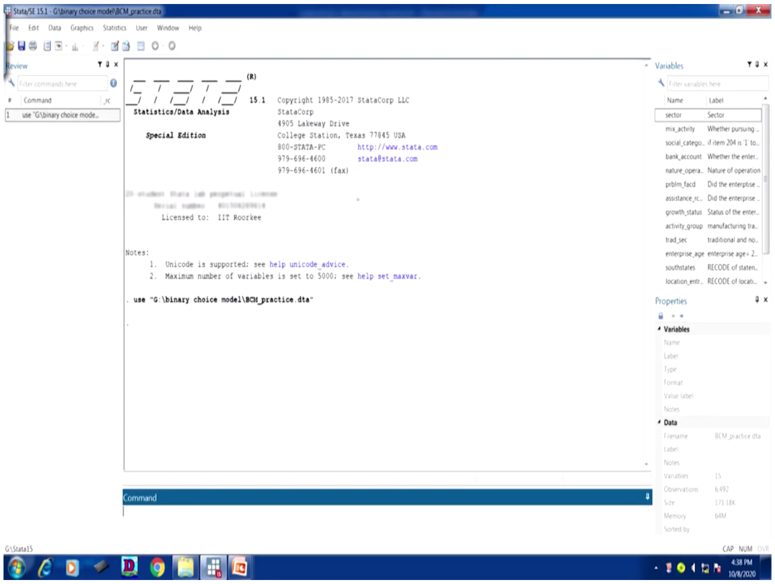


The example we are going to cite is through our data. That is we are providing the sample data to you from 73rd NSS data. So, we have mentioned the title of the data as BCM, binary choice model for practice. And same variables we are linking with the LPM as well as for other Logit and Probit. We are going to compare.

So, running the same model what factors explain the choices of women being necessity entrepreneur. That we already started discussing in the LPM. We are again discussing the Logit function as well. So, which factor explain the choice of women being a necessity entrepreneur. So, that is what we are going to mention.

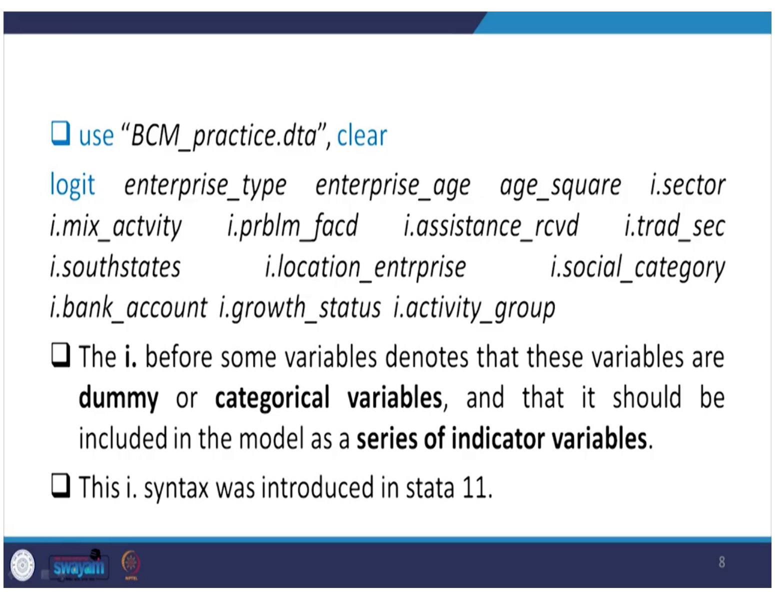
So, we are not summarizing or describing the data as we have already done it in the previous lecture and we have seen broadly that the data consists of bytes not in string; so in numeric, numbers. It is composed of categorical and non categorical as well as in continuous series. But the dependent variable is in binary. And that we have already seen it.

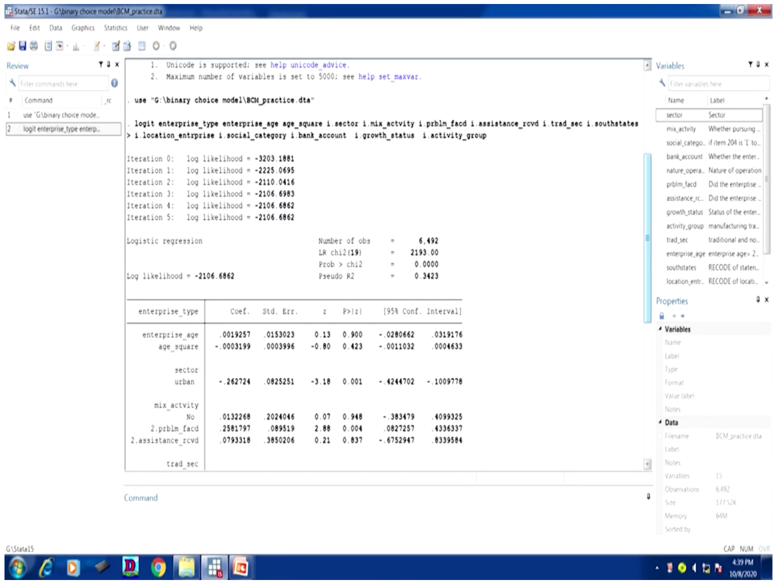
(Refer Slide Time: 13:26)



So, straightaway we are going to run the model. Once we use the model in the STATA, let me open the STATA for you quickly. I am going to open the BCM data. So, it is here, so I am just going to open the data BCM practice data. I told you already. , this has already been opened on the screen.

(Refer Slide Time: 13:47)



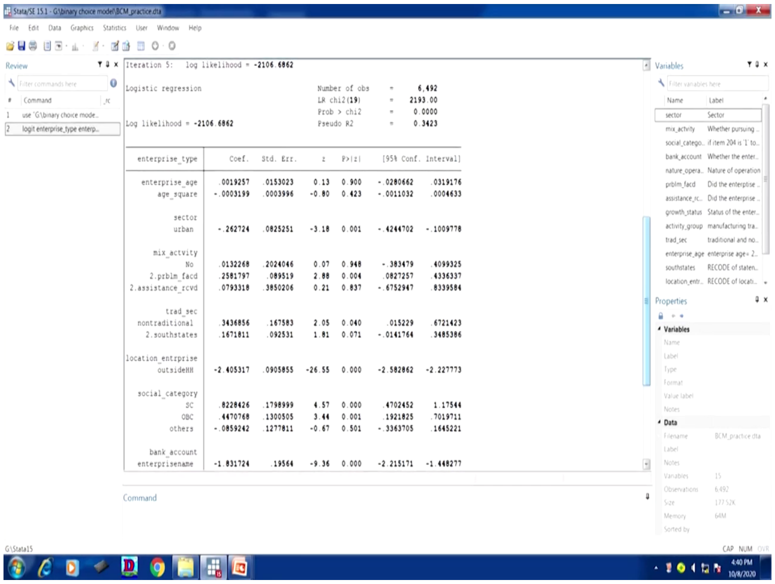


Now, from the data we will now first operate the simple Logit function, the Logit model. So, that Logit model we will estimate quickly. So, it is here. Let me operate all the variables that has been listed and we are running it. So, once I ran all those variables with Logit enterprise, this gives result very quickly. The Logit function has generated the values in front of us, the coefficient values in front of us.

As we already mentioned about R square and regarding pseudo R square, the coefficient already been highlighted, the Chi square value whether that is significant or not, even earlier models we did mention. Once again it is significantly defined. So, most importantly the likelihood function as I mentioned, this was taken 5 iterations. It kept on iterating all its variable with different forms of sample within the model by observing how they are varying.

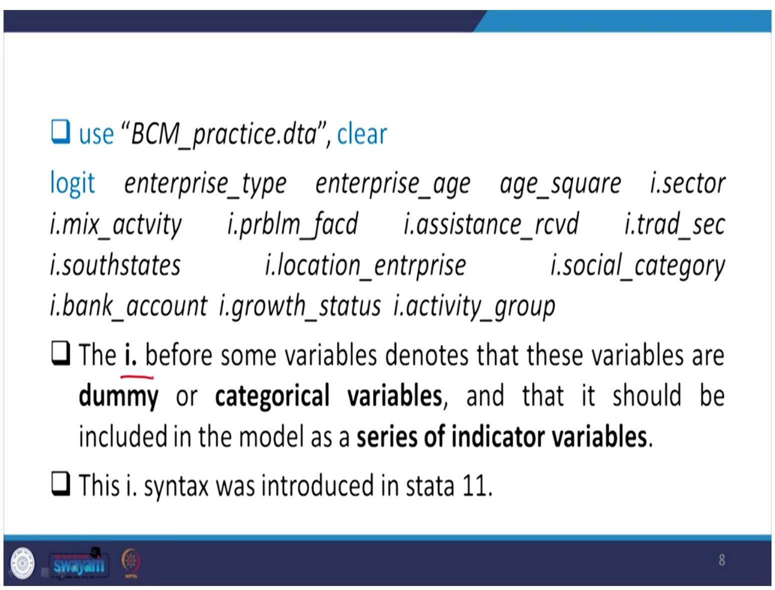
At the end it has iterated with the likelihood value of minus 2106 but this is important to note here whether the module is fit or not, module is fine or not? Whether we go with the model? Because the probability of the significance level of that Chi square value is 0.0000. So, it is not going to be problematic. It rests on which variable are important and how we are going to interpret.

(Refer Slide Time: 15:45)



From the model itself coefficients are given. I have said in the features itself that in Logit, even in Probit also, the coefficient is not important. The odds ratio is important and the marginal effect is very important.

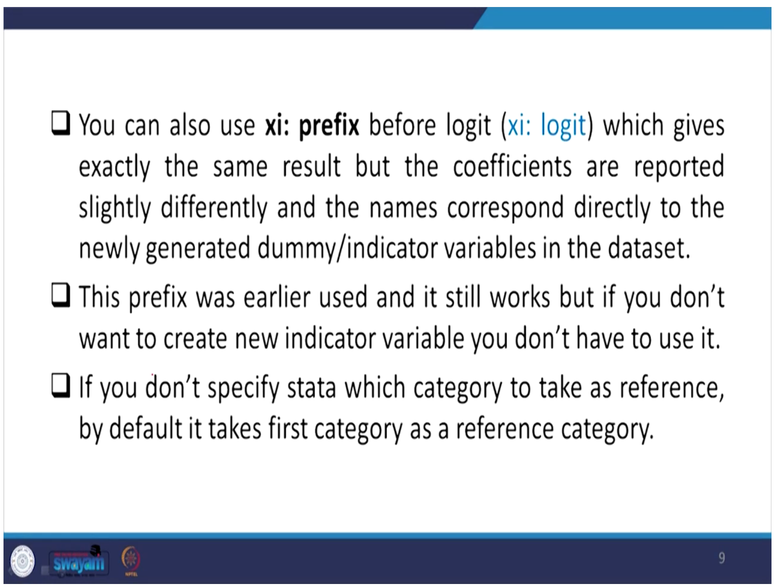
(Refer Slide Time: 16:02)

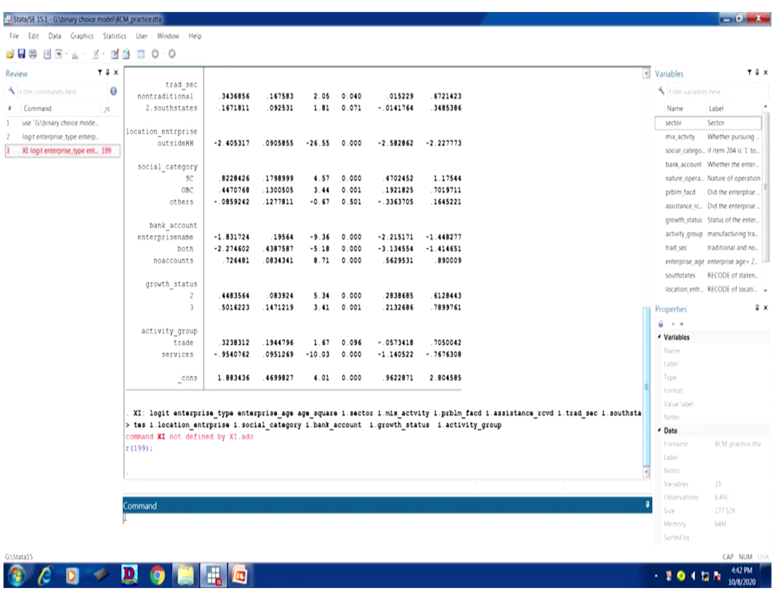


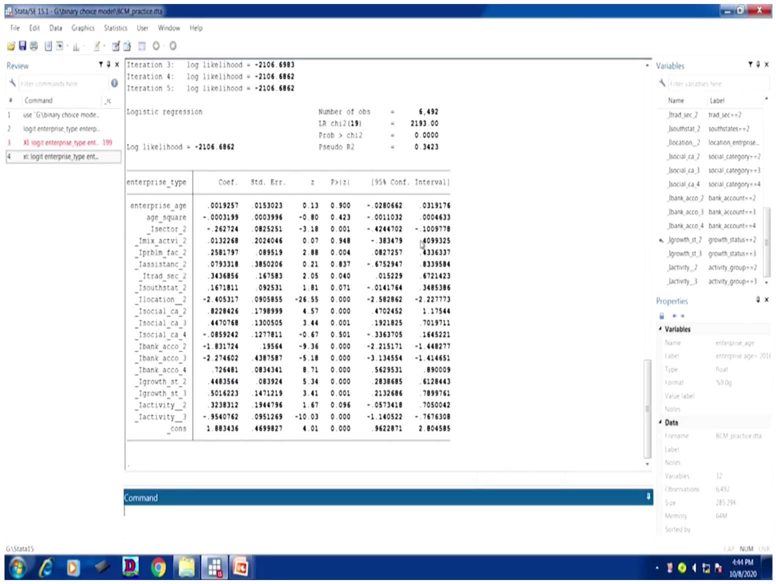
So, we are going to discuss that right now. After doing so as you can see very clearly some of the variables are categorical and some of the variables are dummy. So, dummy variables we already defined in our previous lectures. If you simply take the dummy as an average without having any specification of the dummy, then average of dummy you will not have any correct interpretation.

So, for that you have to take a base value in the dummy. Like to identify the base value we have gone through a command that is this i dot, alright. So, i dot before some variable denotes that these variables are the dummy or categorical, that it should be included in the model as a series of indicator variables, series of indicator variables. So, this i dot syntax was introduced in STATA 11, alright, 11 onwards this has been introduced.

(Refer Slide Time: 17:06)







Those who have earlier version, you might be in trouble. So, you have to take help command for the patching of any extra version of it or extra operated solution to it. We can use xi also, xi from the very beginning, that conditioning for the model with this categorcal variable. So, like prefixing xi logit which gives exactly the same result but the coefficients are reported slightly different and the names corresponding directly to newly generated dummy or indicator variable in the data set.

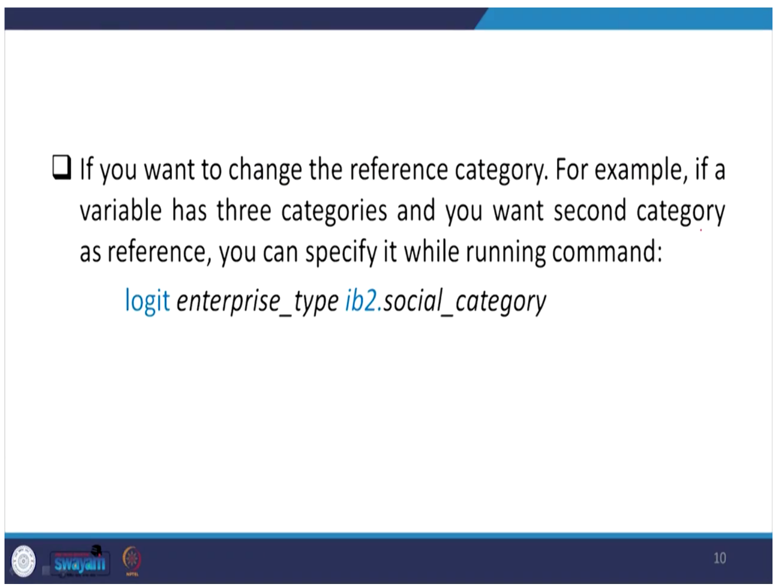
So, we will just show it. Like if you just go by this command, xi command we can go for it. That we can show it here also in the STATA directly. With the same command it is there. This xi colon, if we add xi colon we will get. The look is different. It is in capital. You can make it in small. So, the capital entry as we already mentioned that STATA is very case sensitive so you have to take small letters.

Look at the coefficients, here are interpreted against the exact non-reference category. But in the previous case we just see, look at this. The reference category value is also mentioned or not? Let me see. Only, so even i dot command we have taken. again in the sector, rural and urban so urban value is presented.

Look at this. In the mixed activity also only three are presented. here only one. Each, against each, where is it? In the sector, only the category 2, that category itself is mentioned. It is nowhere mentioning the base category, isn’t it? So, that is the only difference. Suppose you want to get small table I think this is more useful, alright. So, these are also generated as well and defining the other categories as against to the base.

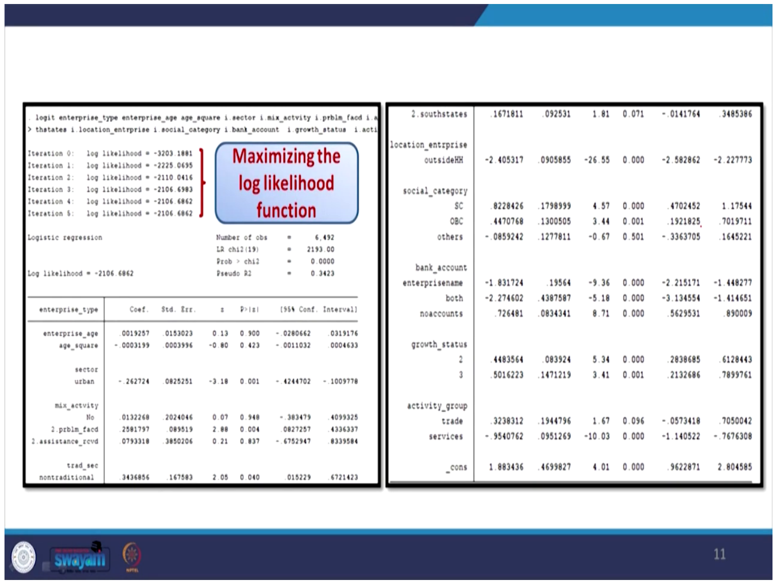
Now, let us go by the detailing here. Coming to the interpretation further, this prefix was earlier used and it still works. But if you do not want to create new indicator variable, basically in the variables itself it identifies the indicator variables, which are the indicator variables, and these are added in the dataset itself. But in the earlier case it is not showing the indicator variables or not added at all. If you do not specify STATA which category to take as reference, by default it takes the first category as the reference category.

(Refer Slide Time: 20:53)



But there are techniques. If we add, it gives the particular category as the base category. Like in rural or urban by default rural was 1 for us, 2 for urban. So, 1 considers as the base category by default. If you specify ib2 dot, ib2 that means it specifies the second one as the base. So, if you want to change the reference category you can accordingly specify. Rest of the details are the same.

(Refer Slide Time: 21:05)



```
. logit enterprise_type enterprise_age age_square i.sector i.mix_actvity i.prblm_facd i.a
> thstates i.location_entrprise i.social_category i.bank_account  i.growth_status  i.acti

Iteration 0:   log likelihood = -3203.1881
Iteration 1:   log likelihood = -2225.0695
Iteration 2:   log likelihood = -2110.0416
Iteration 3:   log likelihood = -2106.6983
Iteration 4:   log likelihood = -2106.6862
Iteration 5:   log likelihood = -2106.6862

Logistic regression                             Number of obs     =      6,492
                                                LR chi2(19)       =    2193.00
                                                Prob > chi2       =     0.0000
Log likelihood = -2106.6862                     Pseudo R2         =     0.3423
```

Maximizing the log likelihood function

enterprise_type	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
enterprise_age	.0019257	.0153023	0.13	0.900	-.0280662	.0319176
age_square	-.0003199	.0003996	-0.80	0.423	-.0011032	.0004633
sector						
urban	-.262724	.0825251	-3.18	0.001	-.4244702	-.1009778
mix_actvity						
No	.0132268	.2024046	0.07	0.948	-.383479	.4099325
2.prblm_facd	.2581797	.089519	2.88	0.004	.0827257	.4336337
2.assistance_rcvd	.0793318	.3850206	0.21	0.837	-.6752947	.8339584
trad_sec						
nontraditional	.3436856	.167583	2.05	0.040	.015229	.6721423
2.southstates	.1671811	.092531	1.81	0.071	-.0141764	.3485386
location_entrprise						
outsideHH	-2.405317	.0905855	-26.55	0.000	-2.582862	-2.227773
social_category						
SC	.8228426	.1798999	4.57	0.000	.4702452	1.17544
OBC	.4470768	.1300505	3.44	0.001	.1921825	.7019711
others	-.0859242	.1277811	-0.67	0.501	-.3363705	.1645221
bank_account						
enterprisename	-1.831724	.19564	-9.36	0.000	-2.215171	-1.448277
both	-2.274602	.4387587	-5.18	0.000	-3.134554	-1.414651
noaccounts	.726481	.0834341	8.71	0.000	.5629531	.890009
growth_status						
2	.4483564	.083924	5.34	0.000	.2838685	.6128443
3	.5016223	.1471219	3.41	0.001	.2132686	.7899761
activity_group						
trade	.3238312	.1944796	1.67	0.096	-.0573418	.7050042
services	-.9540762	.0951269	-10.03	0.000	-1.140522	-.7676308
_cons	1.883436	.4699827	4.01	0.000	.9622871	2.804585

11

- In the result, we can see that the iteration log likelihood indicating how quickly the model converged. The log likelihood is used to compare nested models.
- The likelihood ratio chi-square of 189.65 with a p-value 0.0000 explains that our model as a whole fits significantly better than a model with no predictors.
- Like other regression model, the logit regression also displays coefficients, their standard errors, the z-statistics , associated p-value and the 95% confidence interval of the coefficients.

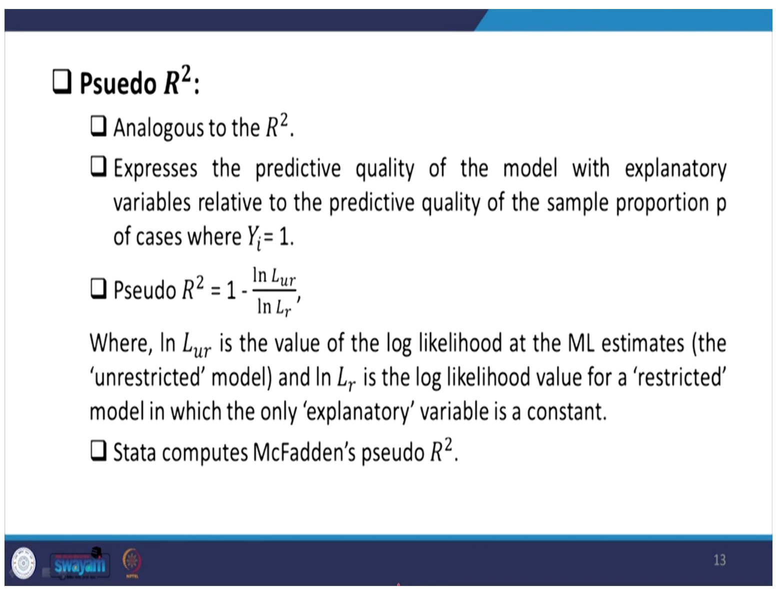
12

We are going to discuss also, we have already discussed about the maximizing of the log likelihood functions, so 5 iterations we already discussed and some other interpretation we already shown to you. So, in the result that we can see that the alteration of log likelihood indicating how quickly the model converged. The log likelihood is used to compare the nested models. If you have any other nested models that likelihood estimator can also be useful for comparison.

The likelihood ratio with Chi square value, we have already shown you Chi square value of 189.65 with a P value that is 0.000. It indicates significant at one portion level explain that our model as a whole fits significantly better than a model with no predictors. So, like other regression model the Logit regression also displays coefficients. So, coefficients that I mentioned. Their standard errors also, standard error next to their coefficients, standard errors are there.

Then the Z statistics also, then the probability limit, the P values, that we can find out which variable is significant. Like in urban it is significant at 1 percent level. These two are not significant. First two are not significant. Then sector urban is significant. In the mixed activities also, the middle one, that is problem faced is significant. Similarly you can interpret other coefficients. So, the confidence interval values are also defined.

(Refer Slide Time: 23:01)

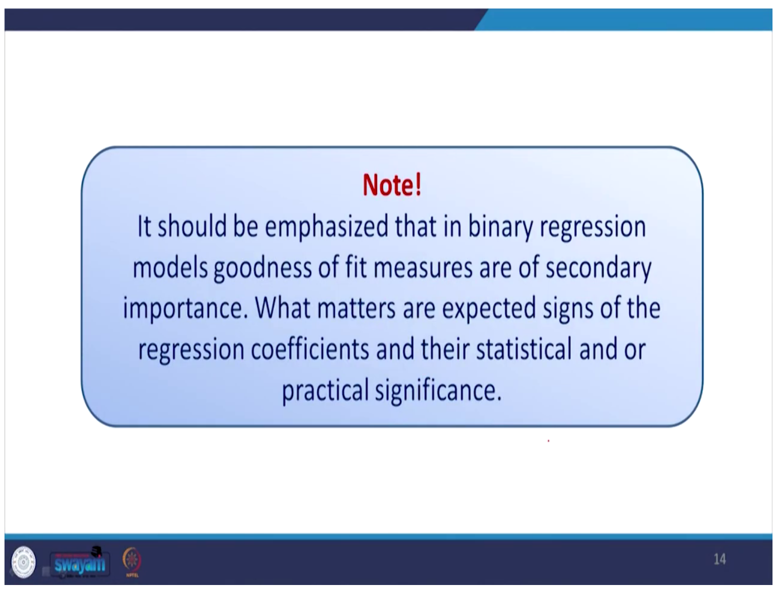


Coming to the interpretation of R square in the model and specifically we clarified in our previous lecture about pseudo R square again, I am emphasizing on the pseudo R square that it is not similar to the R square values we used to have in the continuous case or in the regression model, normal or ordinary least square method but this pseudo R square is analogous to the R square we studied. This expresses the predictive quality of the model with explanatory variables relative to the predictive quality of the sample proportion of cases where Y is equal to success and that is 1.

So, pseudo R square in this case using the likelihood ratio basically takes the log likelihood ratio and with unadjusted or unrestricted and the restricted model. In this case L_{ur} so the pseudo R square is equal to 1 minus log likelihood with unrestricted and log of restricted. In the numerator it is unrestricted.

So, L_{ur} is the value of log likelihood at the ML estimates, that is unrestricted model and the log restricted, R is the log likelihood value of the restricted model which the only explanatory variable is a constant. So, that is important. You can take a note of it. So, STATA computes, already mentioned the McFadden technique of estimating R square that is also called pseudo R square in this case STATA adopts in this kind of model.

(Refer Slide Time: 24:56)



It should be emphasized that in binary regression models, goodness of fit measures are of secondary importance. What matters are the expected sign. The sign, the relationship is more important. Sign of the coefficient is more important and their practical and statistical significance with which we have already discussed is more important than that of R square. So, if R square value is low you may not worry especially in these kinds of models.

(Refer Slide Time: 25:26)

- **Interpretation of coefficients:** Among sectors, as compare to rural, women entrepreneurs in urban sector decreases the log odds of choosing necessity entrepreneurship by 0.262.
- The coefficients of the logit function is quite difficult to interpret since it follows a logistic distribution function.
- As a results we compute the **odds ratio** and the **marginal effects.**
- Stata does this task by specifying simple command:

"logistic" command outputs odds ratios instead of log odds. Use logistic command instead of logit command.

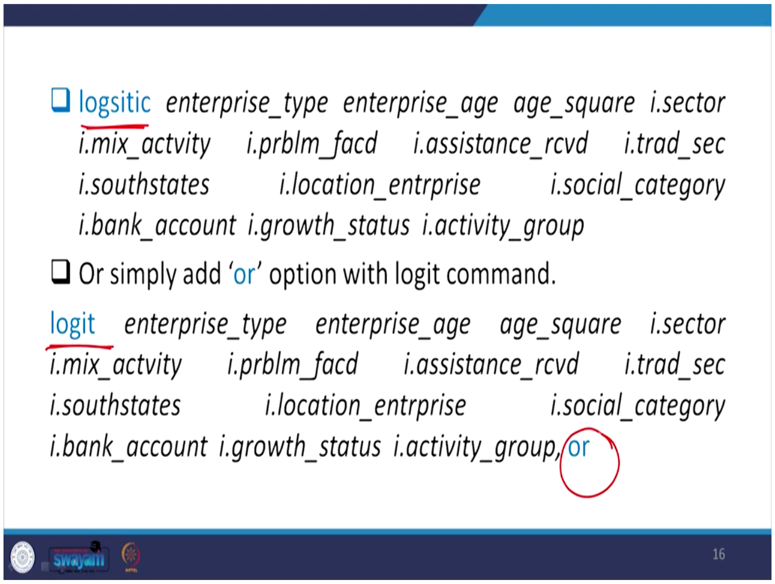

```
. logit enterprise_type enterprise_age age_square i.sector i.mix_actvity i.prblm_facd i.a
> thstates i.location_entrprise i.social_category i.bank_account  i.growth_status  i.acti

Iteration 0:   log likelihood = -3203.1881
Iteration 1:   log likelihood = -2225.0695
Iteration 2:   log likelihood = -2110.0416
Iteration 3:   log likelihood = -2106.6983
Iteration 4:   log likelihood = -2106.6862
Iteration 5:   log likelihood = -2106.6862
```

Maximizing the log likelihood function

```
Logistic regression                             Number of obs     =      6,492
                                                LR chi2(19)       =    2193.00
                                                Prob > chi2       =     0.0000
Log likelihood = -2106.6862                     Pseudo R2         =     0.3423
```

enterprise_type	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
enterprise_age	.0019257	.0153023	0.13	0.900	-.0280662	.0319176
age_square	-.0003199	.0003996	-0.80	0.423	-.0011032	.0004633
sector						
urban	-.262724	.0825251	-3.18	0.001	-.4244702	-.1009778
mix_actvity						
No	.0132268	.2024046	0.07	0.948	-.383479	.4099325
2.prblm_facd	.2581797	.089519	2.88	0.004	.0827257	.4336337
2.assistance_rcvd	.0793318	.3850206	0.21	0.837	-.6752947	.8339584
trad_sec						
nontraditional	.3436856	.167583	2.05	0.040	.015229	.6721423
2.southstates	.1671811	.092531	1.81	0.071	-.0141764	.3485386
location_entrprise						
outsideHH	-2.405317	.0905855	-26.55	0.000	-2.582862	-2.227773
social_category						
SC	.8228426	.1798999	4.57	0.000	.4702452	1.17544
OBC	.4470768	.1300505	3.44	0.001	.1921825	.7019711
others	-.0859242	.1277811	-0.67	0.501	-.3363705	.1645221
bank_account						
enterprisename	-1.831724	.19564	-9.36	0.000	-2.215171	-1.448277
both	-2.274602	.4387587	-5.18	0.000	-3.134554	-1.414651
noaccounts	.726481	.0834341	8.71	0.000	.5629531	.890009
growth_status						
2	.4483564	.083924	5.34	0.000	.2838685	.6128443
3	.5016223	.1471219	3.41	0.001	.2132686	.7899761
activity_group						
trade	.3238312	.1944796	1.67	0.096	-.0573418	.7050042
services	-.9540762	.0951269	-10.03	0.000	-1.140522	-.7676308
_cons	1.883436	.4699827	4.01	0.000	.9622871	2.804585

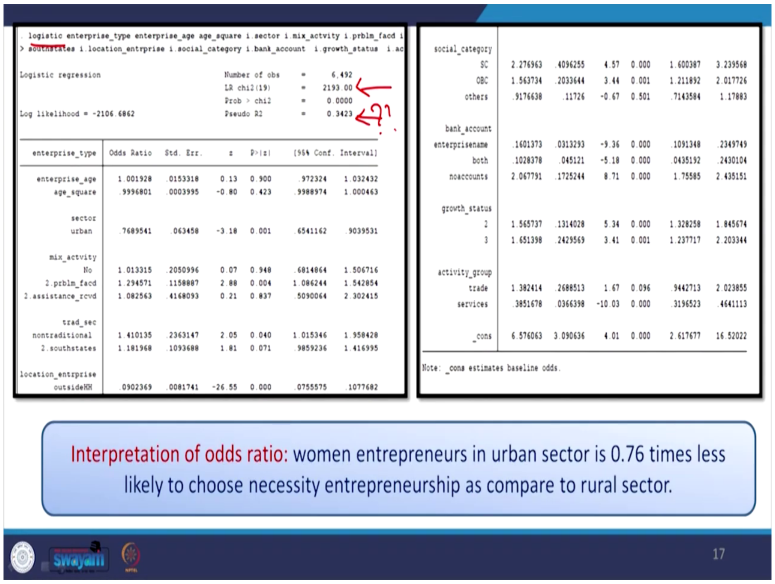


Coming to the interpretation of coefficients further among sectors as compared to rural women entrepreneurs and urban sector decrease the log of odds of choosing necessity entrepreneurs by 0.262. So, like we have already said here. Here it is 0 point in this case. The log of odds is negative related to the necessity entrepreneurs. That means enterprise or the women entrepreneurs to be necessity entrepreneurs being in urban area is negatively linked with a coefficient of minus, coefficient of 0.262, 0.263 roughly.

Similarly, the coefficient of the Logit function is quite difficult to interpret since it follows a logistic distribution function. So, the coefficient which we have said, it is somewhere linking to average. Average coefficient is not suggested to be interpreted. Since the function is logistic it has to be in favor towards success or failure or probabilistic structures is attached. So, likelihood of estimation is attached. So, we should not interpret the coefficient as such. Rather it is suggested to interpret the odd ratio and the marginal effect.

So, STATA does this task by specifying simple command logistic instead of just logit. And so the command is here. We can just operate. Basically whatever we operated it is more or less same but it gives only odds ratio. So, odds ratio logistic command if you simply add here that gives, but if you only logit, till this if you do not add this, it is not going to give odds ratio at all. It only gives coefficient. But if you simply comma add with 'or', or stands for odds ratio. This certainly boils down to the odds ratio of it.

(Refer Slide Time: 27:45)



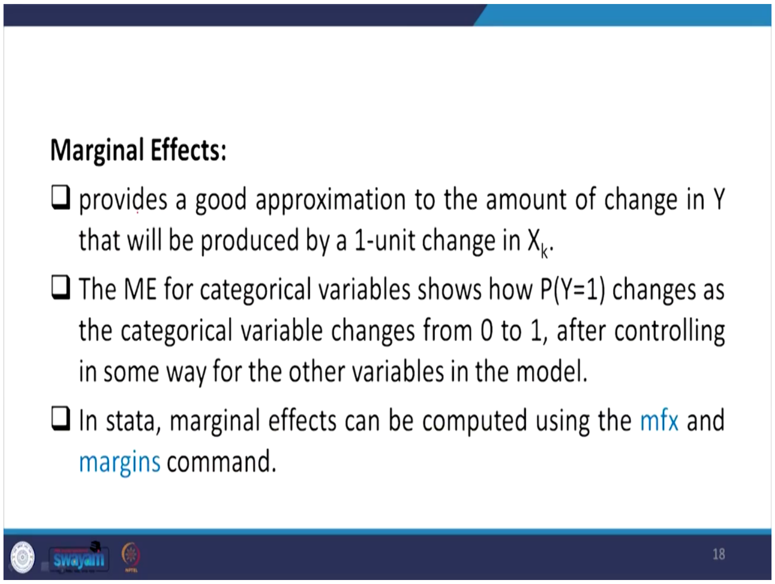
enterprise_type	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
enterprise_age	1.001928	.0153318	0.13	0.900	.972324	1.032432
age_square	.9996801	.0003995	-0.80	0.423	.9988974	1.000463
sector						
urban	.7689541	.063458	-3.18	0.001	.6541162	.9039531
mix_actvity						
No	1.013315	.2050996	0.07	0.948	.6814864	1.506716
2.prblm_facd	1.294571	.1158887	2.88	0.004	1.086244	1.542854
2.assistance_rcvd	1.082563	.4168093	0.21	0.837	.5090064	2.302415
trad_sec						
nontraditional	1.410135	.2363147	2.05	0.040	1.015346	1.958428
2.southstates	1.181968	.1093688	1.81	0.071	.9859236	1.416995
location_entrprise						
outsideHH	.0902369	.0081741	-26.55	0.000	.0755575	.1077682
social_category						
SC	2.276963	.4096255	4.57	0.000	1.600387	3.239568
OBC	1.563734	.2033644	3.44	0.001	1.211892	2.017726
others	.9176638	.11726	-0.67	0.501	.7143584	1.17883
bank_account						
enterprisename	.1601373	.0313293	-9.36	0.000	.1091348	.2349749
both	.1028378	.045121	-5.18	0.000	.0435192	.2430104
noaccounts	2.067791	.1725244	8.71	0.000	1.75585	2.435151
growth_status						
2	1.565737	.1314028	5.34	0.000	1.328258	1.845674
3	1.651398	.2429569	3.41	0.001	1.237717	2.203344
activity_group						
trade	1.382414	.2688513	1.67	0.096	.9442713	2.023855
services	.3851678	.0366398	-10.03	0.000	.3196523	.4641113
_cons	6.576063	3.090636	4.01	0.000	2.617677	16.52022

The results are like this, the way we have shown earlier but I think this is more or less repetitive. You should not operate much. You can operate on your own. The interpretation is like earlier we have seen coefficients but now it is not coefficients. The command here is logistic, more importantly you have to check this first. Then this is not that important. You may not be worried about this pseudo R square but still here it seems 34 percent fitting to the model. It is a good fitting.

Then coming to the interpretation of the coefficient. Wherever it is exceeding 1 that means those variables are positively linked to the dependent variable or the regressand. And those who are less than 1, 1 we are taking as stands for probability of success and so whichever are less they are negative related, whichever are less than 1 they are negative related to the regressand. So, like here these variables, we have already seen that was of minus of 0.262.

But here now look at this. This is negatively linked because as I told you since these are in odd ratio the probability in favor of success or failure, and this is saying that being urban women entrepreneurs is having a likelihood of minus, negatively linked to being a necessity entrepreneur. So, likelihood of the woman to be necessity entrepreneur is negative. The interpretation exactly is that it is 0.7 times less likely to choose or to be a necessity entrepreneur, so being in urban area, alright. Similarly you can interpret other variables, whichever is more than 1 that is, it is positively and otherwise it is negatively linked.

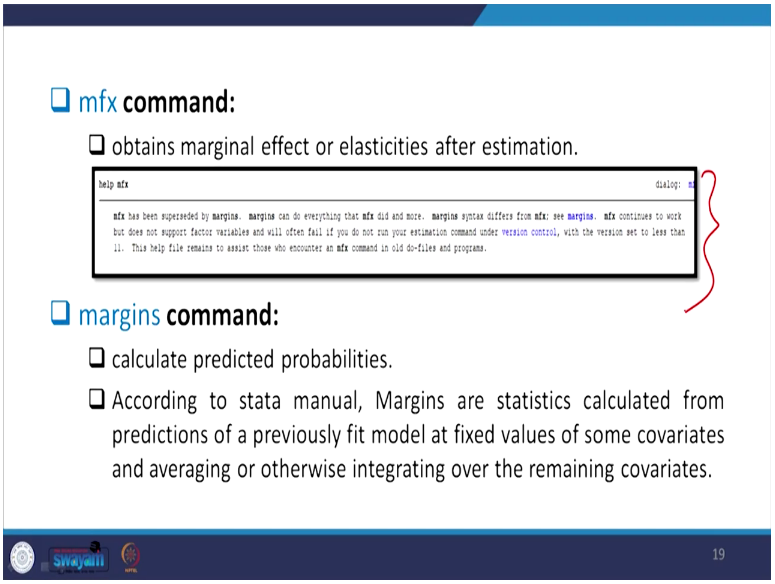
(Refer Slide Time: 30:07)

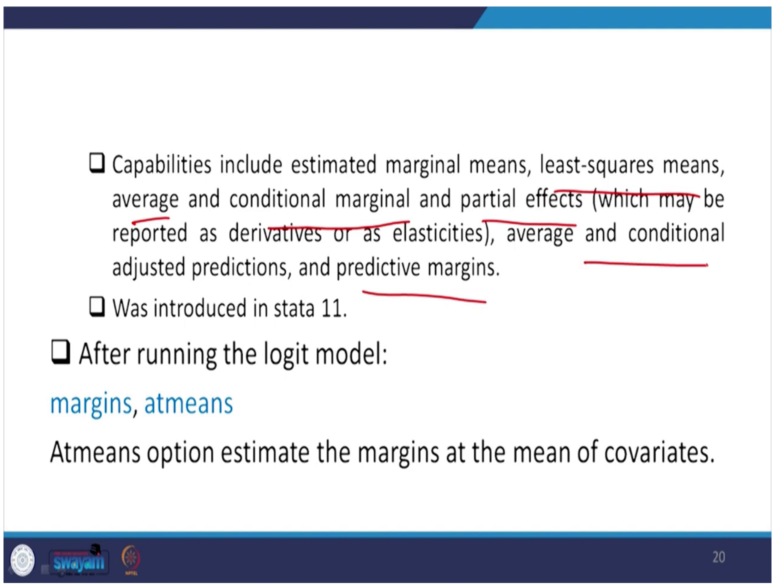


So, marginal effect is very very important. I suggest many times that coefficients are not to be interpreted, better to avoid that. So, straight away come to the interpret of marginal effect. This provides a good approximation to the amount of change in Y given the unit change in X. We have already defined this derivative. So, derivative values of Y with respect to X that the marginal effect depends upon change in X. So, it is important.

The ME that is marginal effect so for categorical variables shows that how the P with value 1 changes as categorical variable changes from 0 to 1, after controlling in some way or the other variables in the model. In STATA marginal effects can be computed using a mfx or the margins command.

(Refer Slide Time: 31:00)



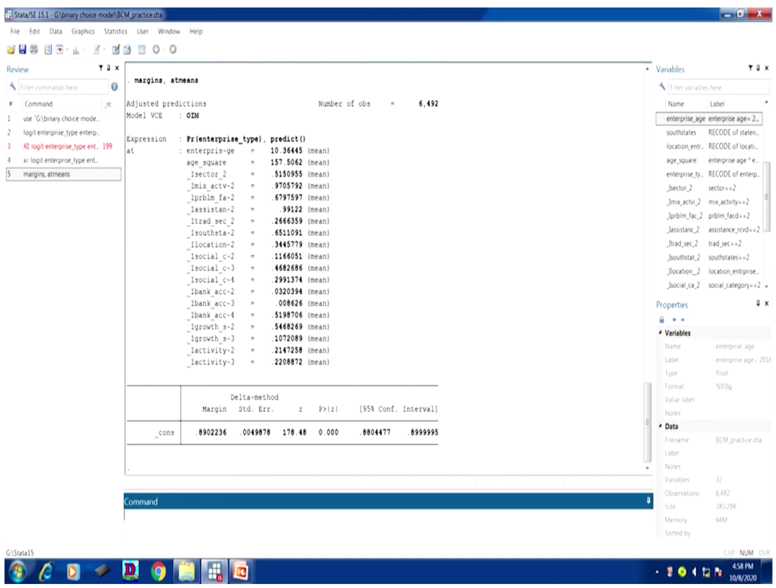


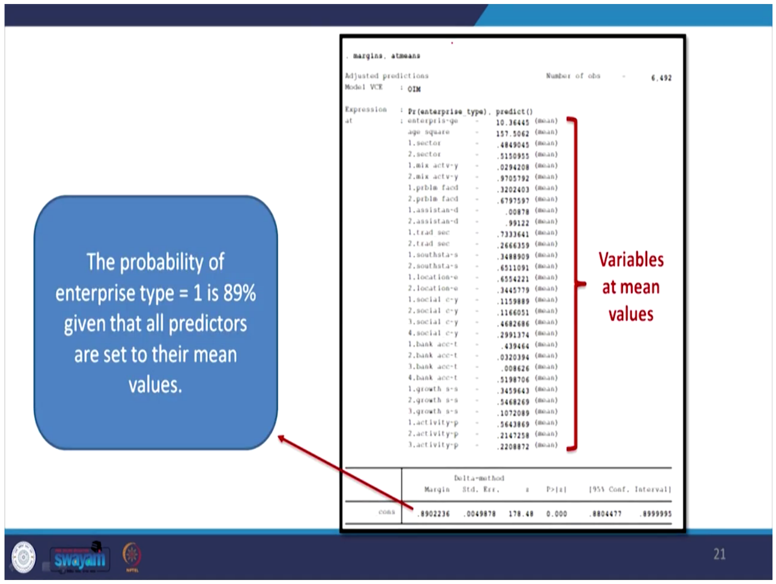
So, you can take the help search mfx command from the STATA. So, many suggestions will be there. You can get it .Out of that we are going to operate with any of the command. We are going to operate with the margins command. So, since we have taken dy/dx so somewhere we are linking to elasticity of their estimation. So, in margins command also helps in giving the predicted probabilities.

According to STATA manual margins are statistics calculated from predictions of a previously fit model at fixed values of some covariates and averaging or otherwise integrating over remaining variables. Basically it is calculated from the predictions of the fitted variables or values. The capabilities of the marginal effect include estimated marginal means. Like it is estimated at means value, then least square means also, alright, average, then conditional marginal, partial effects. There are so many way of interpretation of the marginal values, so which may be also reported as derivatives or elasticities.

Average and conditional adjusted predictions and predictive margins. There are so many ways of interpreting the MEs. So, this was introduced as I mentioned since STATA 11 version. After running the Logit, once we run the Logit model then we can go for margins at means one of the way of interpreting ME. So, atmeans option estimates the margins at the means of the covariates.

(Refer Slide Time: 33:20)



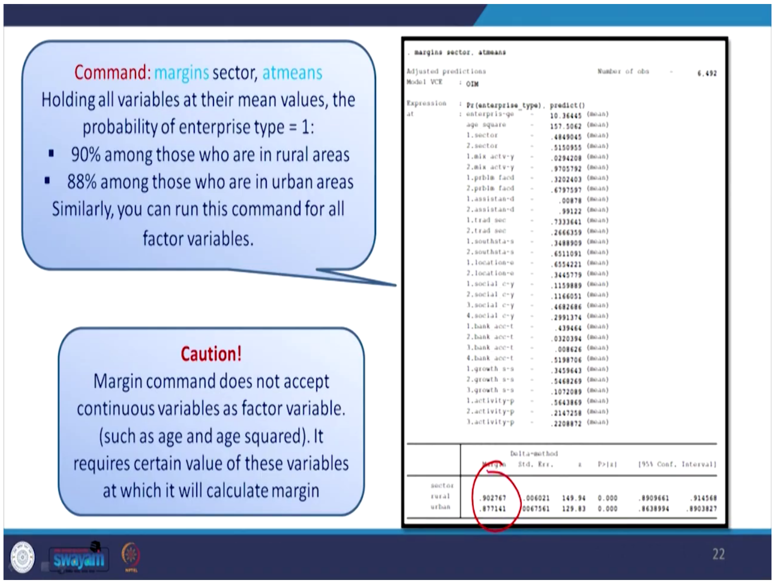


That we just wanted to show you, alright. So, it is here. We simply get that to the model, to the STATA. Since we have already derived the result, we simply take atmeans value. So, margin atmeans, then we go to the STATA window. Then since STATA as already we run the Logit regression. This has given the result.

At the means, margin at its average or the average value or means is derived. So, I am going to discuss it so 0.89 is very very important. And also that predicted values of each of these entries are also equally important for our interpretation. Likewise this is what we derive.

Here the probability of the enterprise with code 1 that is necessity enterprise is of 89 which we have shown that is 89 percent. Given that all the predictors are set to their mean values. All the predictors it is highlighted here variables at means values. You can just have a check. Variables at mean values and everywhere mean is mentioned. So, when all the variables are set at its mean what is the probability of enterprise to be of necessity type? That is of 89; nearly 90 percent is having the probability of to be necessity driven or necessity enterprises.

(Refer Slide Time: 34:57)



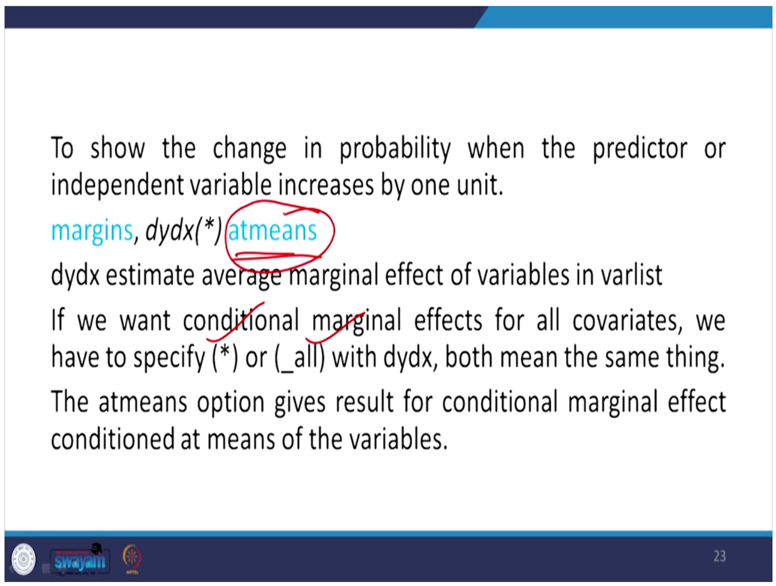
Coming to specific command emphasizing sector and their margin, particular sector-wise, which sector has higher probability, rural or urban? If you specify with margins with a sector with means. Initially we said only margins atmeans that gives the total model, isn't it? When we specify to a particular categorical variable, we just wanted to know difference between rural and urban probabilities. It has given the value. The result we can easily get with the same command if you can try.

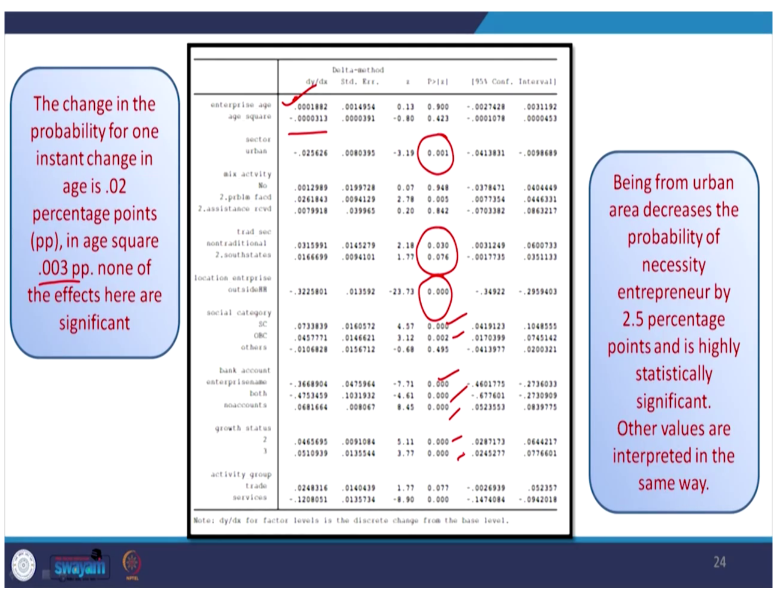
For the rural and urban; rural it has given 90.902767. So, that means 90 percent probability are there with the rural entrepreneur to be necessity driven or necessity type of enterprises whereas in the urban areas of 87. So, even it is understood and logical that in urban areas the likelihood of the enterprises to be necessity driven is very less, should be less than that of the rural areas because of many other problems or facilities or shortage of access specially financial access as well, transport and financial access as well.

So, coming to the understanding relating to the margin command this does not accept continuous variable. That is more important because we are comparing with categories. So, this does not accept continuous variable as a factor variable such as age or age square is not going to be interpreted at all or is not going to be considered. It requires certain values of these variables at which it will calculate the margin. It will compare from one to another one. There must be some discrete difference from one category to another category then only the margin could be derived. Then only the derivative could be derived.

So, let us come to the understanding of the change in the probability. What are the change in the probability when the predictor of the independent variable increases by 1 unit? So, if there is, we simply say the absolute probability of a particular category when we keep other variables to be at their average value.

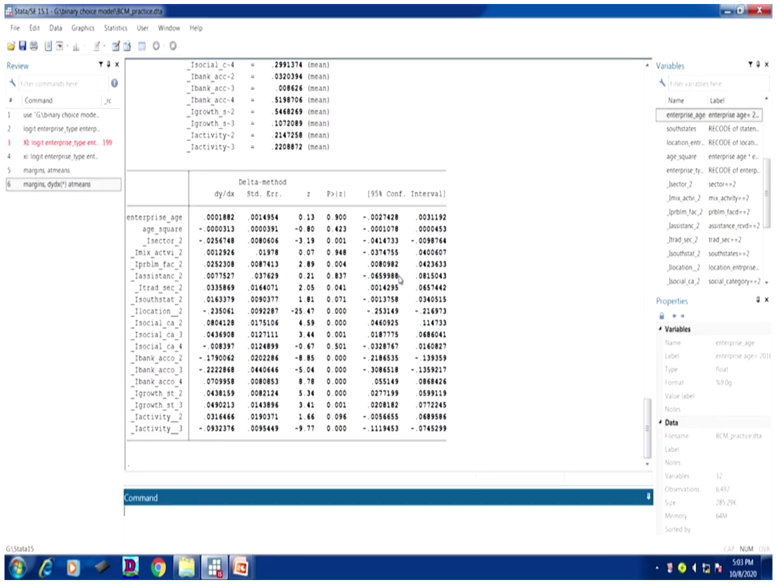
(Refer Slide Time: 37:15)





	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
enterprise_age	.0001882	.0014954	0.13	0.900	-.0027428	.0031192
age_square	-.0000313	.0000391	-0.80	0.423	-.0001078	.0000453
sector						
urban	-.025626	.0080395	-3.19	0.001	-.0413831	-.0098689
mix_actvity						
No	.0012989	.0199728	0.07	0.948	-.0378471	.0404449
2.prblm_facd	.0261843	.0094129	2.78	0.005	.0077354	.0446331
2.assistance_rcvd	.0079918	.039965	0.20	0.842	-.0703382	.0863217
trad_sec						
nontraditional	.0315991	.0145279	2.18	0.030	.0031249	.0600733
2.southstates	.0166699	.0094101	1.77	0.076	-.0017735	.0351133
location_entrprise						
outsideHH	-.3225801	.013592	-23.73	0.000	-.34922	-.2959403
social_category						
SC	.0733839	.0160572	4.57	0.000	.0419123	.1048555
OBC	.0457771	.0146621	3.12	0.002	.0170399	.0745142
others	-.0106828	.0156712	-0.68	0.495	-.0413977	.0200321
bank_account						
enterprisename	-.3668904	.0475964	-7.71	0.000	-.4601775	-.2736033
both	-.4753459	.1031932	-4.61	0.000	-.677601	-.2730909
noaccounts	.0681664	.008067	8.45	0.000	.0523553	.0839775
growth_status						
2	.0465695	.0091084	5.11	0.000	.0287173	.0644217
3	.0510939	.0135544	3.77	0.000	.0245277	.0776601
activity_group						
trade	.0248316	.0140439	1.77	0.077	-.0026939	.052357
services	-.1208051	.0135734	-8.90	0.000	-.1474084	-.0942018

Note: dy/dx for factor levels is the discrete change from the base level.



We are saying, what is their marginal unit change. In that case we have to take the dy/dx. So, dy/dx if you can take it that will be giving result like this. We can find it also. So, let us have a check. So, dy/dx and this is going to be more useful throughout your papers. While you write for assignment or papers it will be very useful.

This has given the marginal effect which we are trying to interpret. Against against each predictor variable we have now derived this marginal effect, the unit change, the individual derivative of each variable on the dependent variable. So, the marginal change is highlighted and these coefficients should be interpreted. Those were positive, negative very clearly understood. Generally it is highly linked with the coefficient values but the coefficient values if you interpret, that is misleading. So, better stick to this.

dy/dx estimate average marginal effect of variables because we have set at mean values, so average marginal effect is derived. If we want conditional marginal effect for all covariates we have to specify with the star or all with dy/dx, underscore all with dy/dx. Both mean the same thing. Either this or this refer to the same interpretation. So, if you simply go by this that will give you the conditional marginal effects. The atmeans option gives result of conditional marginal effect conditioned at the means of the variables. That is the conditional aspect we are referring to. Then we have to specify it.

In this case the change in the probability of one instant change in particularly age here. Age is basically minus of 0.0003. So, the change in percentage is of 0.02 alright. So, age is here. The first one is age. Second one is age square. When the second one square case is age square that is of here, that is 0.003.The first one is 0.001882 so roughly around 0.02 point.

So, coming to the discussion of urban area as against to the rural, in this case, urban it is negatively linked to the necessity entrepreneur. In percentage point it is of 2.5 percent lesser points. So, being from urban area decrease the probability of necessity entrepreneurship by 2.5 percent and is highly statistically significant. That we wanted to check. This is statistically significant. Similarly, you can understand this is statistical significant. Even these are statistically significant at different levels. Then these variables, I am just mentioning. You can understand all three. Similarly other you can compare.

(Refer Slide Time: 41:03)

Alright, so from the next class we will try to understand Probit and Tobit. I think it is very clearly understood to you that how Logit is operated and how marginal effect at conditional mean is understood and it is important for interpretation. And please go and practice with any other dataset but sticking to the kind of specification with the dependent variable and independent variable. Then that really works better. So, let us explore Probit regression model from the next class. Thank you so much.