Exploring Survey Data on Health Care

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

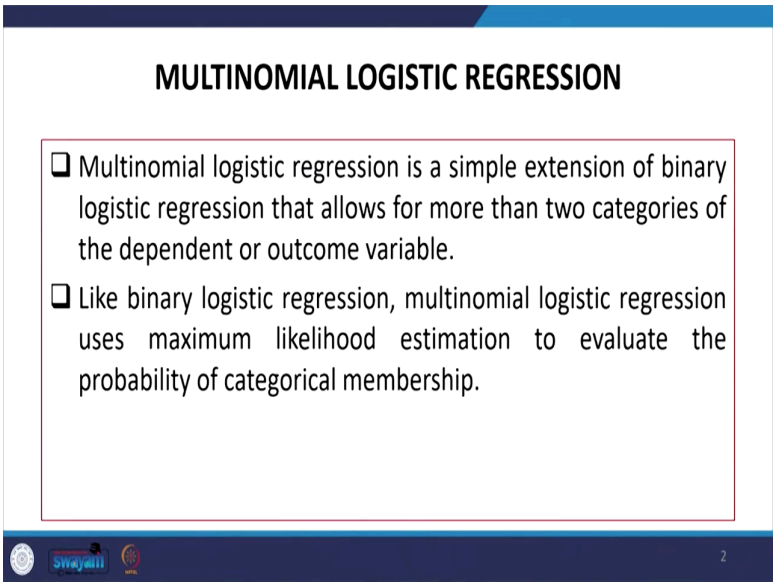
Lecture - 30
Independence of Irrelevant Alternatives

Welcome participants once again to the NPTEL MOOC module on exploring health care data. We are in the 6th week and explaining one of the topics which is sometimes required. Because some reviewers may ask about cross checking or validating your models, i.e., on Independence of Irrelevant Alternatives.

Though it seems very unique very different, but it is connecting to some of the specific models. It is usually discussed in the limited dependent variable models especially after multinomial logit. So, here it goes. Multinomial logistic regression we have already discussed in the previous lectures. We are just trying to understand or validate the multinomial logit/ logistic regression with the help of this particular topic.

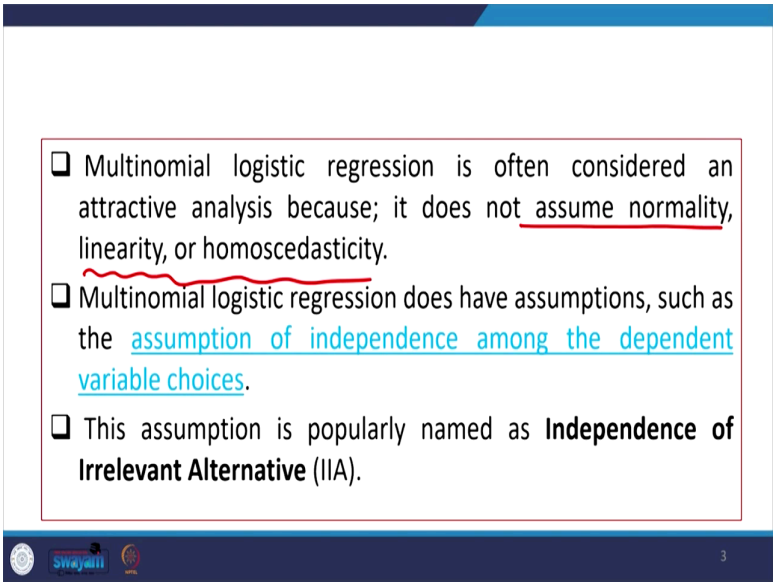
I am just revisiting the multinomial logistic regression once again just for your clarity to introduce the independence of relevant alternatives. Multinomial logistic regression is in fact a simple extension of binary logistic regression i.e., it allows for more than two categories of the independent or outcome variables. So we already discussed in the previous lecture that when we have more than two categories of the dependent or the outcome variable and not the independent one. I think I said as independent. You just correct it as dependent one.

(Refer Slide Time: 02:17)



When the dependent or the outcome variable has more than two categories, we use multinomial logistic regression. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation technique to estimate the probability of categorical membership.

(Refer Slide Time: 02:37)



Now, a multinomial logistic regression is often considered an attractive analysis. Because it does not assume normality or linearity or homoscedasticity. So, these are important aspects which is very essential to discuss at this moment. That is why, these kind of models in case

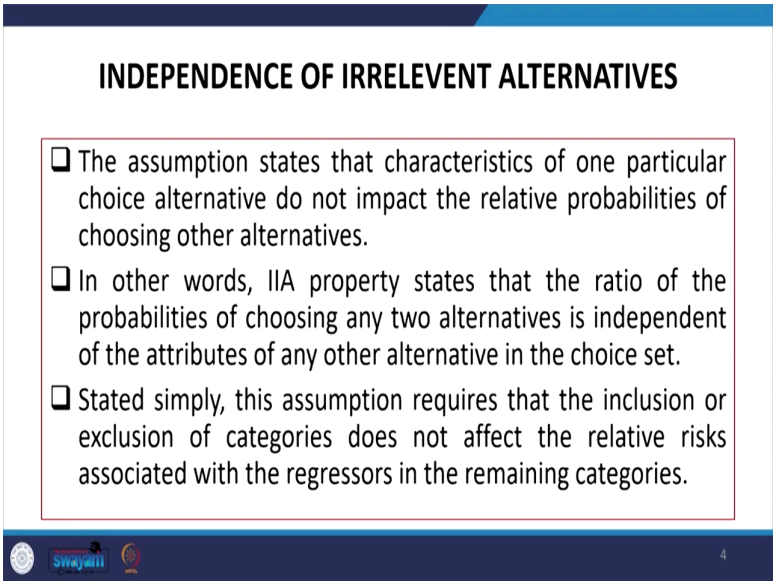
when your data is not following the assumptions of normality or linearity or it does not have homoscedasticity or there are certain problems of heteroscedasticity. Then multinomial logistic regression could be used.

This is also sometimes referred in the context of GLM though there are different context in GLM, but where the linear regression is not applied the GLM can also be discussed generalized linear models. So, multinomial logistic regression does have assumptions such as i.e., called assumption of independence among the dependent variable choices.

The dependent variable choices are the categories which we are sincerely going to consider in the model to run the multinomial logistic regression, though choices should be actually independent. If there are independence among the choices I think if you are including or deleting certain categories or choice in your model, it is not going to be affected much.

This assumption is popularly known as Independence of Irrelevant Alternatives or IIA. So, IIA whenever is discussed, so be confident about it and learn as per we just discussed.

(Refer Slide Time: 04:36)



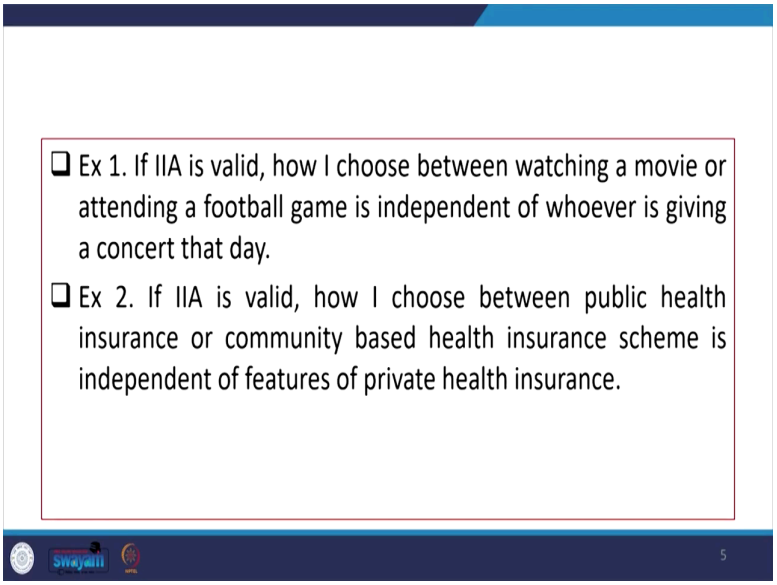
So, what does this mean i.e., Independence of Irrelevant Alternatives? This means that the assumption which IIA just says/states that characteristics of one particular choice alternative do not impact the relative probabilities of choosing other alternatives. In other words, IIA property states that the ratio of the probabilities of choosing any two alternatives is independent of the attributes of any other alternatives in the choice set. This means the ratio

of the probability of choosing any two alternatives is also independent of the attributes of any other alternative in the choice set.

These state simply that this assumption requires that the inclusion or exclusion of categories does not affect the relative risk associated with the regressions in the remaining categories. So, the relative risk like those who are in the context of health care for them this relative risk ratio is also taken largely this is called a multinomial logistic regression.

So, bother change in certain categories or choices the risk is not going to be changed or relative risk is not going to be changed much.

(Refer Slide Time: 05:57)



One example is that of independence of irrelevant alternative is valid. How I choose between watching a movie or attending a football game is independent of whoever is giving a concert that day. So, like if another one we are adding as giving concert.

When I am quite sure that between the choice in movie and attending football is in fact independent of another categories. Another one is if IIA is valid, how I choose between public health insurance or community-based health insurance scheme, is independent of the features pro of private health insurance.

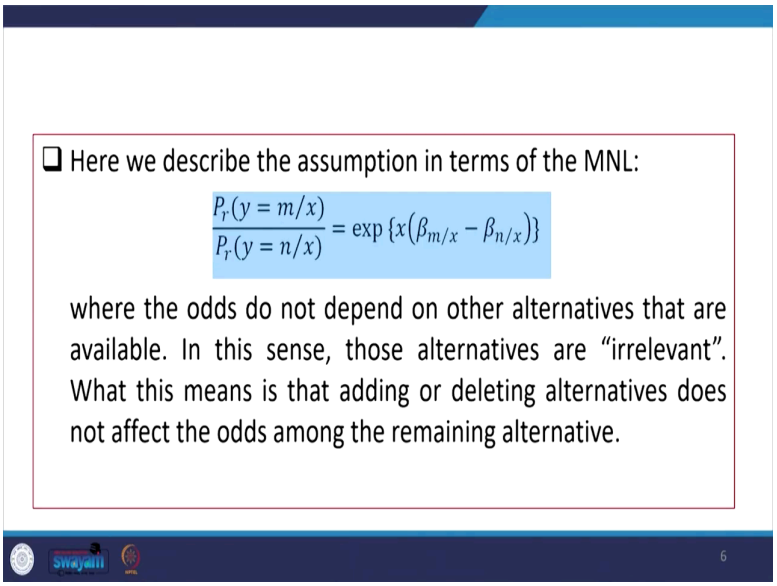
If the choice has already been discussed between these two. And if the third one is introduced that also not creating much difference in the model, that is why it is called independence of

irrelevant alternatives. We discuss and describe the assumption in terms of the multinomial logistic regression/ logistic regression with the help of this equation.

Probability of why conditioning of the variables for, m alternative options, why having alternative categories options with respect to the or it its ratio with respect to other alternatives i.e., still m given its covariates that is x or the control variables in basically an exponential function.

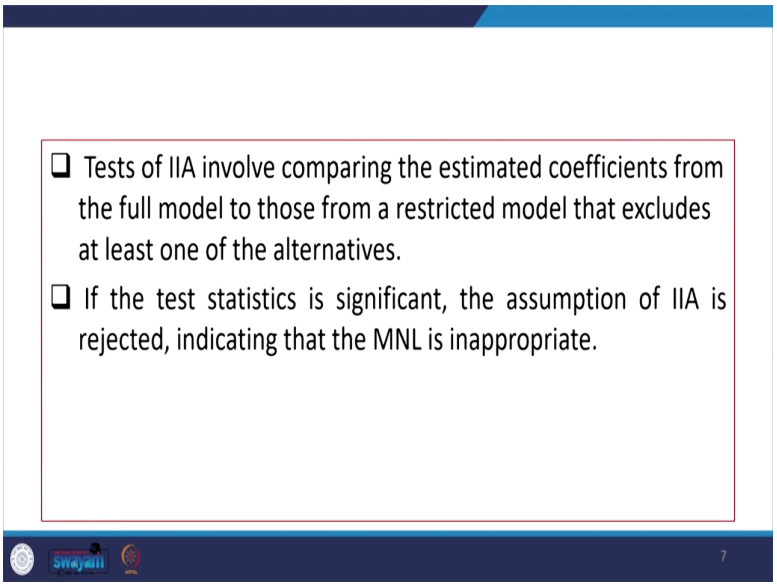
Which we generally know that logistic multinomial logistic functions are expressed in exponential form. That is the exponential form x within bracket beta m given x minus beta n given x.

(Refer Slide Time: 08:09)



Where the odds do not depend on other alternatives that are available in the sense those alternatives are irrelevant. What this means is that adding or deleting alternatives does not affect the odds among the remaining alternatives.

(Refer Slide Time: 08:24)



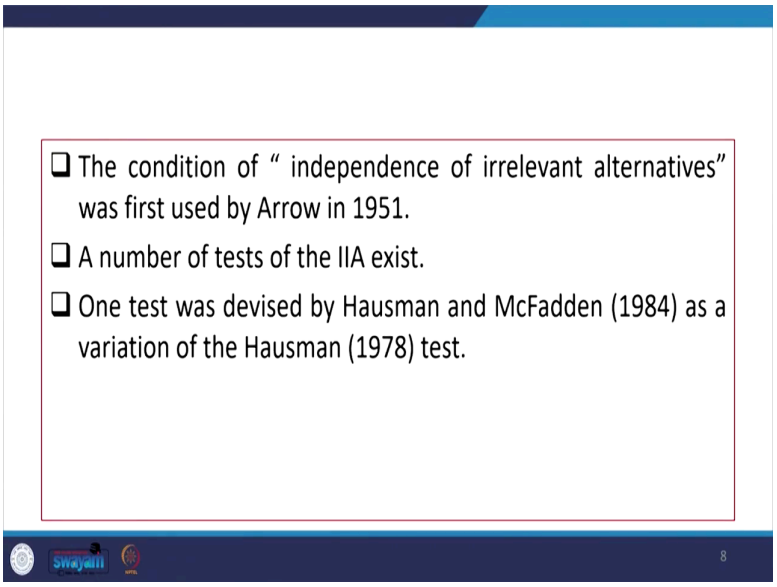
Test of IIA involve comparing the estimated coefficients from the full model to those from a restricted model that excludes at least one of the alternatives. The test which we involve comparing the estimated coefficients from the full model to those from a restricted model restriction model where some alternatives are actually excluded. So, we will compare these two basically in this case, when this comparison by exclusion is not making difference between that of the full mode.

That means, we can assure ourselves that the alternatives are not that important or are independent. If the test statistics is significant; that means, were assumption is there difference now, the assumption of IIA is rejected. Indicating that MNL is inappropriate; that means, there is difference here. we are seeing there are differences. So, like if the test statistics is significant the assumption of IIA is rejected.

If there are any alternative assumption is there that is rejected this means the MNL which you have taken is in fact. I mean the assumption of independenceness is no longer there when we are rejecting; that means, test (Refer Time: 09:48) significant. So, there are in fact dependence not independence of the alternatives, when there are dependents then if you are including certain choices limited choices in the multinomial logistics regression that may not be perfectly explaining the complete model.

So, our we are going to explain you about the test statistics in our practical session here.

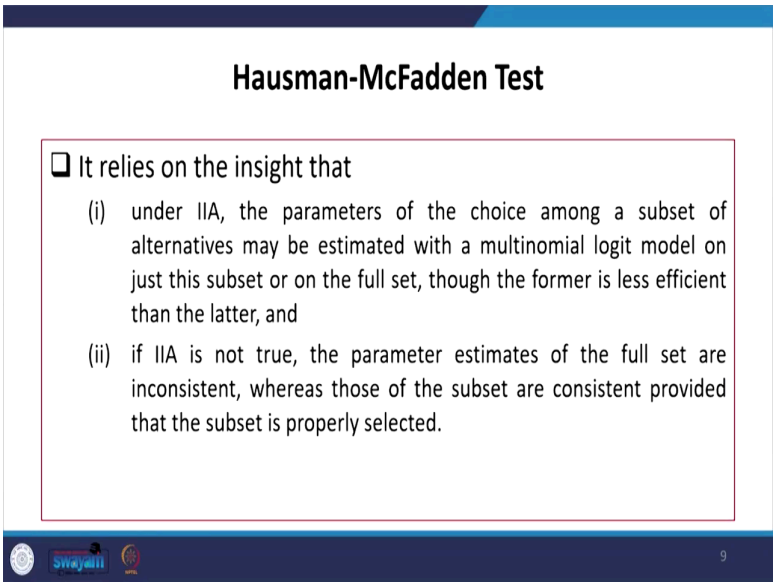
(Refer Slide Time: 10:16)



The condition of independence of irrelevant alternatives was first used by arrow in 1951. A number of tests of I independence of irrelevant alternatives exist. One test was devised by a Hausman McFadden this is what we are going to discuss. As for the 1984 paper as a variation of the Hausman 1978 test.

So Hausman McFadden test, this is relied on the insight that under IIA the parameters of the choice among the subset of alternatives may be estimated with a multinomial logic model, on just this subset or on the full set though the former is less efficient than the latter.

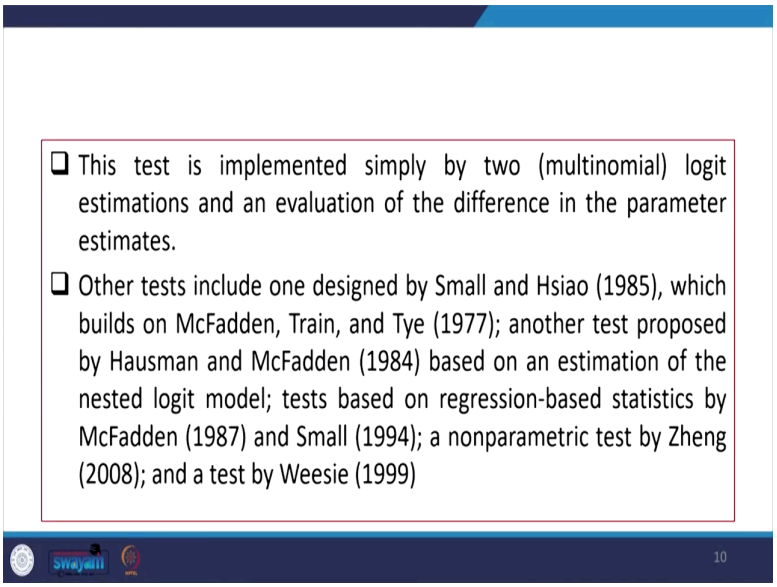
(Refer Slide Time: 11:00)



Because since we did not have restricted our model first one is the subset and second one is the full set, since we have restricted our model. So, that is not expected to be more efficient than that of the full model. If IIA is not true; that means, if there is no independence of alternatives are not independent; that means, we have to reject the parameter estimates of the full sets are inconsistent.

Whereas those of the subsets are consistent provided that the subset is properly selected this test is implemented simply by 2 logistic estimations and an evaluation of the difference in the parameter estimates. The difference of the parameter of these two models are most important.

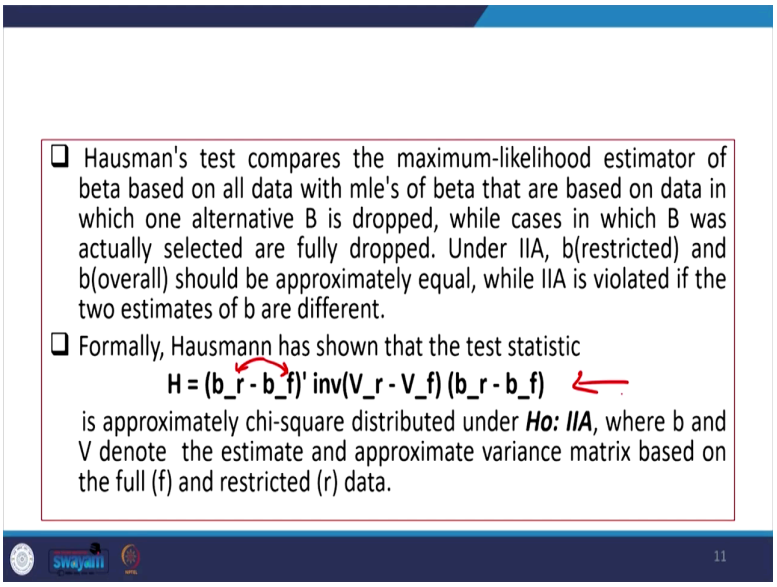
(Refer Slide Time: 11:58)



Other test includes one designed by Small and Hsiao in 1985 paper which builds on McFadden, Train and Tye paper. Another test proposed by Hausman McFadden 1994, based on the estimation of the nested logit model. Basically nested with the alternatives included or not.

Then test based on regression based statistics by McFadden as well in 1987 and Small 1994 and non parameter test model was also suggested by Zheng and a test by Weesie 1999.

(Refer Slide Time: 12:36)



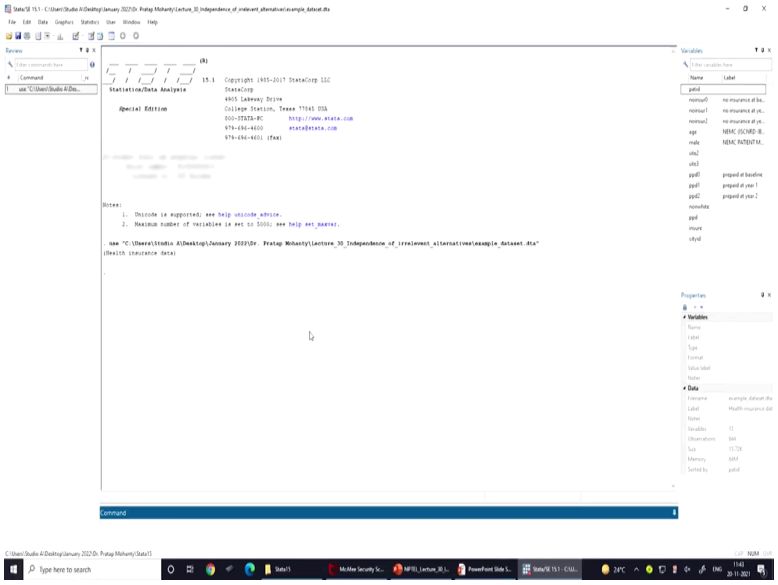
Hausman test compares the maximum likelihood estimator of beta coefficient based on all data with maximum likelihood estimations of beta. That are based on data in which while one alternative beta is dropped while cases in which beta was actually selected are fully dropped. Under IIA beta restricted and the beta overall should be approximately equal.

If they are approximately equal; that means, there is no difference while equal, while IIA is violated if the two estimates of beta are different. If these two estimates of beta are different then IIA is in fact violated. So, in the test that should not be significant if it is significant that mean there are differences.

Formerly Hausman has shown that the beta statistic is like this H is equal to b r minus b f inverse of this V r and V f and its difference is multiplied. This is approximately as the chi square distributed under the null hypothesis as independence of all alternative hypothesis where b and V denote the estimate and the approximate variance matrix based on full f and the restricted. r stands for the restricted model and f stands for the full model.

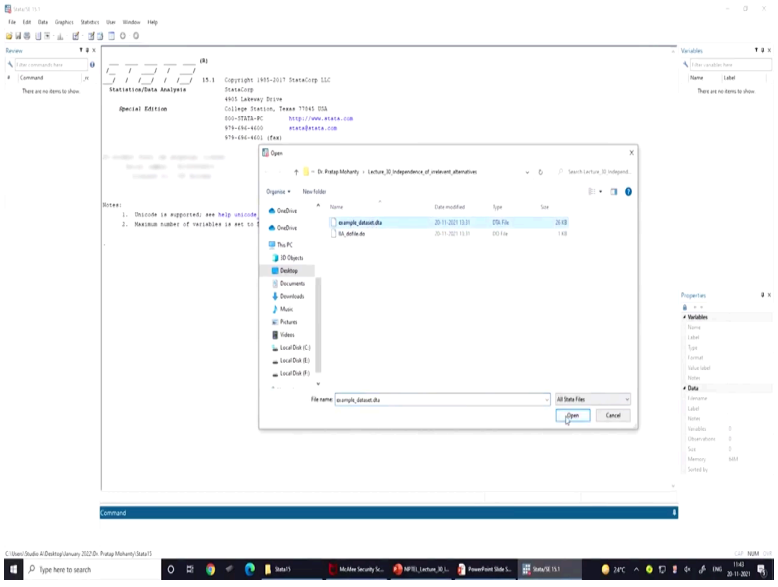
So, this difference we are just getting the difference between these two and in each of the cases. One is in the case is b for the estimate and V for the variants. So, each indicator is included in this equation. So, in the session we are just going to clarify through our approach.

(Refer Slide Time: 14:35)



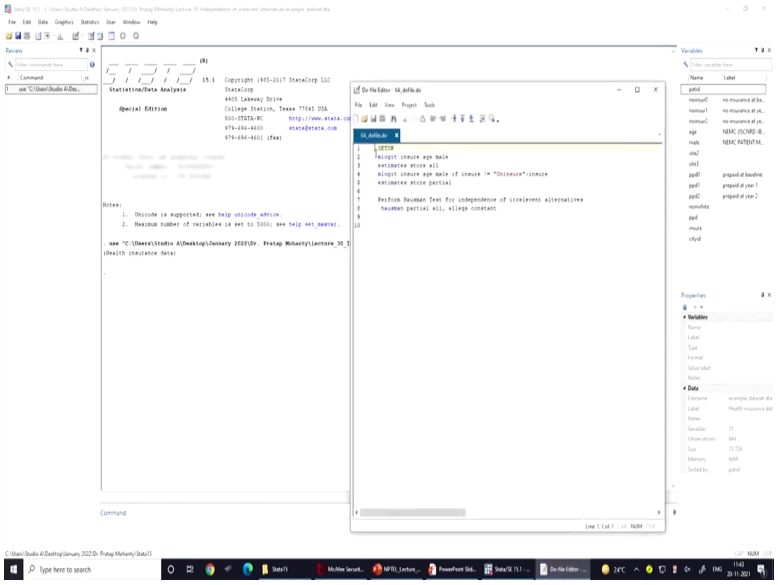
So, here are the details in on your screen this is what we have opened STATA data window on the screen.

(Refer Slide Time: 14:41)

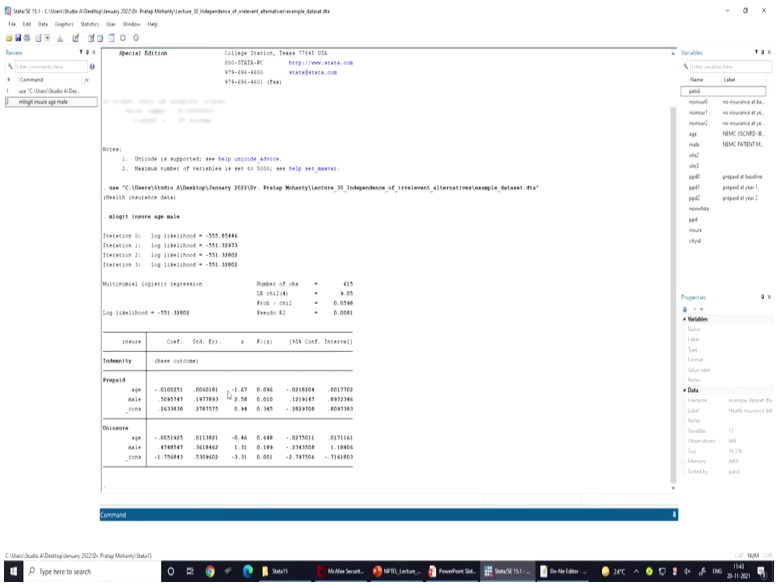


Now, we are going to open the data set which we have sorted out a sample data set. We are also going to share with you. And this is based on insurance which you have already discussed that for two insurance if the access is there and the third choice if you are including and if it is not making much difference. That means, there is independence of irrelevant alternatives.

(Refer Slide Time: 15:07)



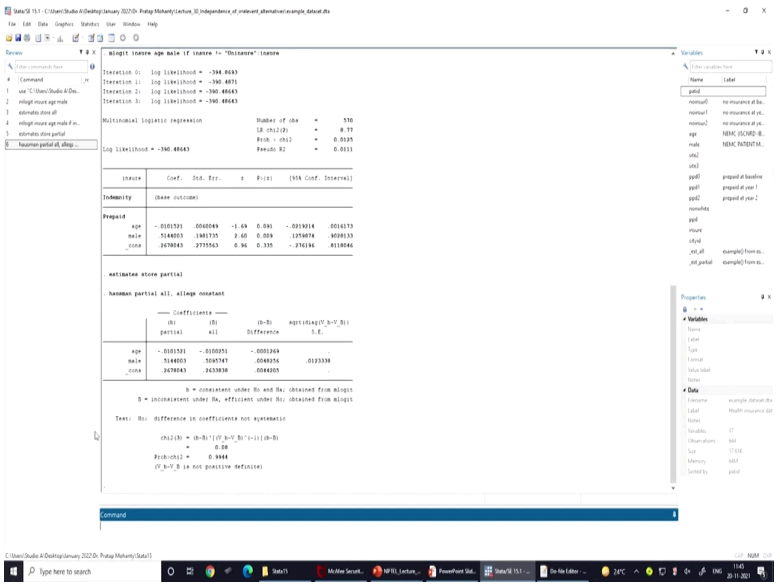
(Refer Slide Time: 15:26)



So, now we are going through by do file. We will also suggest what are the commands. Like first one is we will run the multinomial logistic using two variables. Insure is in fact our dependent variable. And this is our first model alright. m logit we have derived, then we will go for its estimates and then we will store all its estimates.

This is the second command, we usually give third one is we will compare this on insure to insure. In the m logit command this is the third command we have to give.

(Refer Slide Time: 15:59)



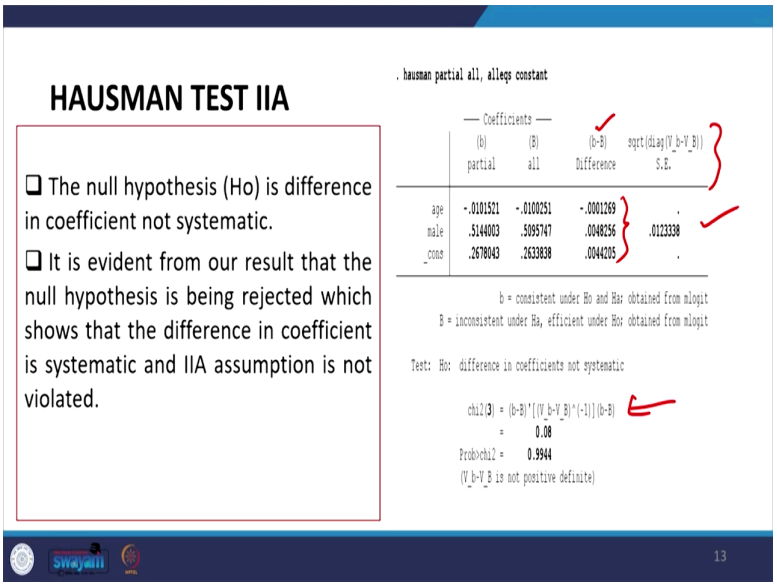
So, now we have got to these differences, the last one we wanted to check with their estimates and their partial values as well. So, then we will perform after having these differences in the m logit we will finally, perform the Hausman test. So, to do Hausman test of irrelevant alternatives. So, this is what our chi square value which you have already said this is what the equation displayed on your screen.

b minus capital B and its variants b stands for the r that is restricted model small b capital B for the full model and its coefficient based on that the chi square value is on your screen. And that is in this case is as discussed is not significant. The coefficient is 0.08 and not significant.

So, if it is not significant which already, we have said if the chi square model based on these suggestions by a Hausman if this comes out to be significant; that means, there are difference between the restricted model and that of the full model.

And in the restricted model we have dropped one category and in the full model all entire options are taken. So, the difference if since it is not significant; that means, there is no difference between these two models even with certain restrictions. So, that does indicate that if you are adding know more variables in the dependent variable category in that case this is not going to differentiate our result. So, that way we say Independence of Irrelevant Alternative holds correct.

(Refer Slide Time: 18:20)



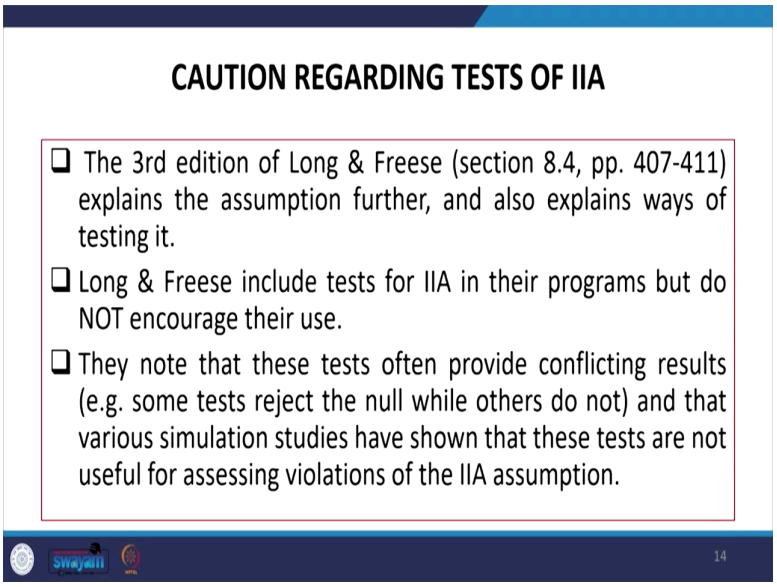
So, this is what we have given in our in our model. In our slides as well Hausman test, this is what is the coefficients we just derived and the difference we have also derived small b minus capital B, capital B stands for the full model.

So, each of the difference is actually derived here. In the square term as, but the suggestions given is also calculated. And then a test is taken, for the assumption here is that there is no difference in coefficient and that coefficients are not systematic, so non systematic assumptions were taken. The null hypothesis is in fact, the is a difference in coefficient that is non systematic.

So, systematic used to be non-random, non systematic is in fact random. So, the difference whichever is derived between these two is expected to be having a distribution that is maybe normally distributed. Since in our model we have said that the difference is itself insignificant. So, there is no distribution of those differences. So, the in the model is in fact following IIA correctly.

It is in evident from our result that the null hypothesis is being rejected which shows that the difference in coefficient is systematic and IIA assumption is not violated.

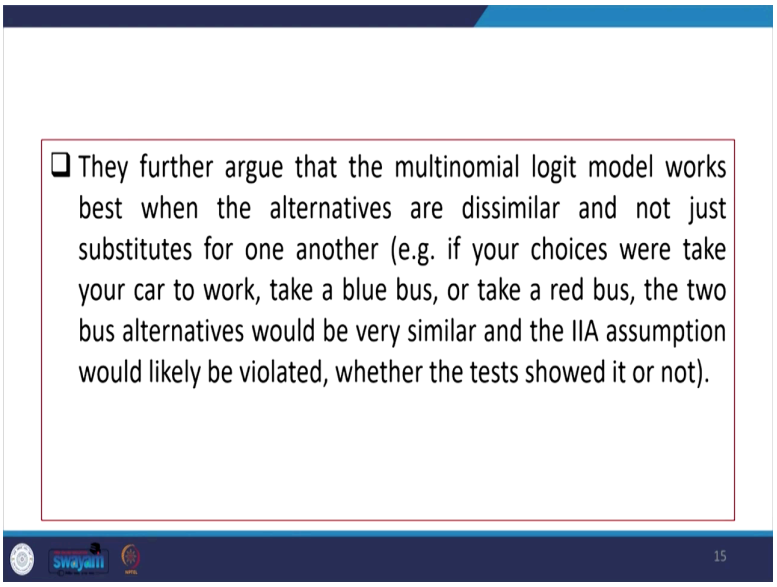
(Refer Slide Time: 19:54)



So, what is this then, like what are the cautions we need to take regarding IIA test some of the caution? The third edition of long and freeze as for the pages and their reference we have cited explains that the assumption further explains about this and also explains the ways of testing it. Long and freeze also test for IIA in their programs, but do not encourage their use. They strongly say that do not use this and do not encourage to use it. They note that these tests often provide conflicting results.

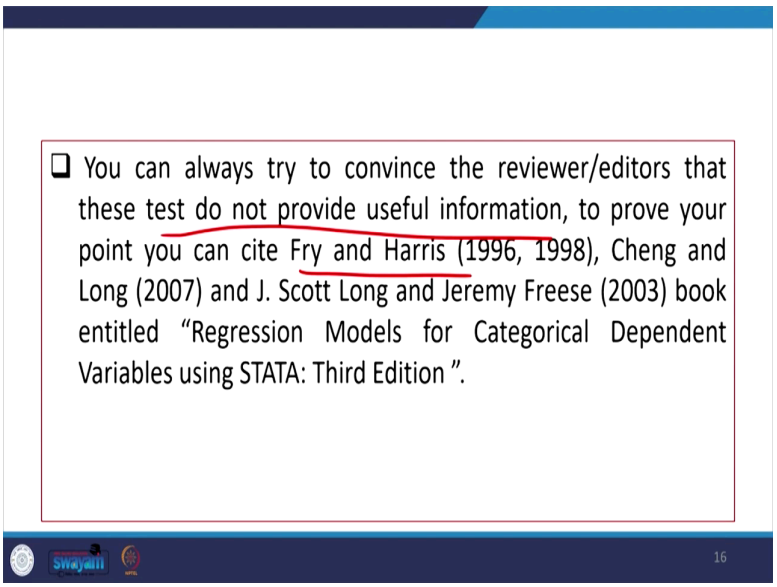
Because some tests reject the null while others do not. And that various simulation studies have shown that these tests are not useful for assessing violations of the IIA assumptions.

(Refer Slide Time: 20:57)



So, this has to be noted carefully that you may avoid using these techniques. They further argue that the multinomial logit-based model works when the alternatives are dissimilar and not just substitutes for one another. That is, if your choices were taken in your car to work or like, whether like take your car to work, take a blue bus, take a red bus. The two bus alternatives would be very similar. And the IIA assumption would likely be violated whether the test sold it or not.

(Refer Slide Time: 21:40)



So, these are the details on another clarification. You can always try to convince the reviewers if you are receiving this type of comment from the reviewers. You just always try to clarify about the IIA as to convince the reviewers. That or editors that these tests do not provide useful information.

To prove your point, you can cite Fry and Harris papers, Cheng and Long Cheng, J. Scott and Jeremy Freese book entitled – "Regression Models for Categorical Dependent Variables using STATA" 3rd edition.

So, in the conclusion though we have explained in this in one lecture after reading thoroughly and we thought that we should give you one direction.

For your clarification some reviewer asks for some clarification about your Independence of irrelevant alternatives. But it is suggested that if run different model it may not be necessarily going to be insignificant every time. And this is not a very robust way of saying the fact that alternatives are not going to differentiate your result. So, it is suggested that you please avoid this kind of result.

But if you just want to give you the minimum information to the reviewer then you can calculate by using the formula or the package given in the do file we have share. That will be very helpful and this is although a very tiny issue, but sometimes useful in research.

With this, I think we should close here and we look for your participation in the next class, for a better lecture.

Thank you.