Multinomial logistic regression, Part 2: Multiple multinomial regression

Author: Heini Väisänen

Transcript of: https://youtu.be/hbdKrRvrEd8

This video is part of an NCRM Online Resource

Hi everyone, my name is Dr Heini Vaisanen and I'm a lecturer in social statistics and demography at the University of Southampton. This video is about multinomial regression and this is the part two multiple multinomial regression, so if you haven't watched the part one yet, please watch that video first before embarking on this one. So the outline of today's presentation is basically that we will look at a multinomial logistic regression model that has more than one explanatory variable and then we will talk about model selection which becomes relevant when we are adding more variables than just one like we did in the first video. So we will look at how we can use likelihood ratio tests and what tests to look at the statistical significance of the variables and how they can help us conclude which variables will keep in our model.

So like I said in the first video we only had one explanatory variable, however, in most applications we would actually want to have more than one in order to control for some other variables that might be associated with our outcome, while examining the effect of the others, and I will show you how to include, interpret and check the statistical significance of these variables in these kind of situations.

As before, we will go through this by using an example. So we will continue looking at economic activity and like in the first video our outcome has three categories, so people are either economically inactive, unemployed or in employment and in the first video we only added gender as an explanatory variable but this time we will also add age and marital status. Before we go ahead with the regression model I will show you some descriptive statistics.

As you might remember from the first video most of the people in our sample, which is coming from an ONS opinion survey while being moduled from April 2011, 56% of the people in the sample were in employment, that was the largest group. About 40% were economically inactive and about 5% were unemployed. When it comes to gender there are slightly more women than men in sample, 54 versus 46%. Here's the distribution of the other variables. So we have categorical age and as you can see we have a respondents ranging from age 16 to 75 plus and the largest age group is 25 to 44 year olds, that's about 35% of the sample and it's the largest group because it's a very wide age range and this is the age range that we would normally expect to be working, whereas the younger ones might still be in education and then some of the, well especially if you go to 65 plus groups then they might be retired. Then we have some information about marital status, more than half of the participants were either married or co-habiting with a partner, about 20% were single and just under 25% where widowed or divorced.

So let's start building our model on the basis of the first model in the first video where we had added gender and nothing else. Let's start by adding age, so we've now added the categorical age and as you can see our reference category is the youngest category 16 to 24 because that is, that doesn't appear on the slides or on the results. These results are from Stata by the way, if you use R or SPSS or some other program it might look slightly different, but the information that you get should still be more or less the same. So here we have the information is presented as ratios rather than log odds which we could also chosen. In Stata if we want to see if age is significant in the model, if we should include age in the model we could start by looking at the statistical significance of the individual categories that you see in the table on the very right hand side, but it's a bit difficult to say based on that, there are some categories that are significant but especially in the economically inactive versus unemployment model and you have actually quite a lot of small p values but then in the unemployed versus unemployment model you actually have quite large p values, so it's difficult to say what exactly is going on with age. So what we can do instead is we can use the likelihood ratio test, to test whether age as a whole is significant in the model. In Stata you would do that by storing the estimates of both models, so the model with only gender and then the model with gender and age, which and you can see the code on the slide estimate store model 2, which is now the second model that we've run and if you go through the computer workshop you will learn how to do this yourself, and then test the differences between model one, which is only gender and model two, which is gender and age. Remember that likelihood trace your tests to two nested models, so models that have some variables in common, but then there is one model that is larger than the first one, so it has more variables, and in this case it's only age that has been added, so our likelihood ratio test tells us whether age is significant and it looks like it is very much so here because we have a very small p-value, so we should include age in the model. When it comes to the effect of age, we can look at the odds ratios that we have here on the slide and I will give you an example from the economically inactive equation. So for instance if you look at the age group 25 to 44, we can see that we have an odds ratio of 0.38 or 0.39 and what that means is that the odds of the age group 25 to 44 of being economically inactive rather than in employment are smaller than those who are in the age group of 16 to 24, and how much smaller, well 61% smaller odds than in the youngest age group.

Then in the second step we add marital status and again we can conduct a likelihood ratio test to figure out whether marital status should be included in the model. Again if we just look at the p-values that we can see on the right hand side column in the table, we have some high p-values and some low p-values and so we want to know where where the marital status, as a whole, is significant in the model and again in the computer workshop materials you will see how to do that yourself, but the result is here and we get a very small small p-value indicating that we should keep marital status in the model.

So now let's look at the association between marital status and unemployment to give you another example of how to interpret the results. So I've taken out some information from the table that we saw that came from Stata, into this more simplified table. So we have the two equations, we have unemployed compared to employed and we have economically inactive compared to employed and then we have only the marital status variable shown, but this is still from the model where we also have gender and age, I've just taken them out from this table to make it a bit simpler to look at. So let's start by looking at the unemployed equation. The odds ratio for the married co-habiting category was 0.18 and our reference category is single. So we can say that when we control for age and gender, married people have 82% lower odds of being unemployed, rather than employed, compared to single people and that this pairwise association is significant because we have a very small p-value. So you might want to comment on that p-value as well as the p-value that I get from the likelihood ratio test, which tests marital status as a whole, and again like you saw in the first video, since we have a dummy variable and we're conducting a multinomial regression it can get a bit complicated with the multiple comparisons, so you might want to calculate some predicted probabilities instead if you want to make the interpretation a bit easier.

If we wanted to say something about the widowed/divorced group in the unemployed equation, we could say by looking at the odds ratio that when we control for age and gender, widowed or divorce people had about 0.3% higher odds of being unemployed rather than employed compared to single people. However, this effect is very small and the pairwise association is not statistically significant, so it looks like those who are widowed or divorced are not statistically significantly different from single people.

Like I said before, it can be more straightforward to use predicted probabilities when we're making interpretation in multinomial models, and you can either do this, you can calculate them by hand but actually usually it's quicker to do it in statistical software, but here are the equations just in case you will end up doing this by hand and also to show you where the predicted probabilities are coming from.

So again we have three categories of the outcome like we had in the first video and if you compare these equations to the ones that I showed you in the first video they look very similar other than the beta1 x1 part and beta2 x2 part of each equation has been bolded. When we have a bold part of an equation like this it basically means that that is a vector of all of the exponential variables that have been included in the model. So it doesn't mean that we would only include one exponential variable when we calculate predicted probabilities, it means that we include all the relevant exponential variables for that combination of characteristics and we calculate that. And again remember that in the last category was the reference category, so that's the category for employed people, and the equation is slightly simpler than for the other two because the numerator is one instead of it being an equation.

So how would we calculate these probabilities? First of all we need to figure out what the log-odd values were for our model and you can see an extract here, I've taken out some information that we don't really need here, because in the earlier slides you only saw odds ratios. If you want your software to do this for you, you could use margins command and Stata which you will see in the computer workshop, but let's do it by hand. So here in this table we have the relevant numbers that we're going to need to plug in to our equations.

So let's say that we want to figure out the probability of someone who is married and in the age group of 25 to 44 and who is a man of being unemployed, inactive or employed. We take the constant or the intercept from the tables that you saw on the previous slide, plug that in that's negative 1.36, then we take for the unemployment equation and then we take the coefficient for being married in the unemployment equation which is negative 1.96, and then we take the coefficient for the relevant age group which is 25 to 44 in our example and the coefficient is negative 0.001. Then we take the other two equations, so the other equation that we have, so that is the inactive equation, plug in the values from that, so we plug in the intercept, the coefficient from rate and the coefficient for age, which are slightly different this time negative 1.15, negative 0.11 and negative 0.89 respectively, and we, for the denominator we have 1 plus the original equation and the inactive equation. That gives us the probability for being unemployed for someone who has, who is a man aged 25 to 44 and who's married, and we do the same thing for the inactive equation. So this time the numerator has the equation that we've taken from the inactive equation in the original model and the denominator stays the same as before and then finally for the probability of being employed we divide one by the same denominator as we had in the other two equations.

If we calculate these, if we solve these equations or if we use statistical software to do this for us, these are the results that we get. So a man who's aged 25 to 44 and married, the probability of that man of being employed is 86% according to our model, a probability of being unemployed is 4% and the probability of being economically inactive is about 10%. I've also calculated the probabilities for the other two marital categories, so for single men and divorced and widowed men, and you can see that there are some differences by marital status in, when it comes to what probability these people have of being employed, unemployed or inactive. So married people are more likely to be employed than single people, or divorced or widowed people, it's 86% compared to 72% in the other two groups. Married people are less likely to be unemployed, so they only had 4% probability of being unemployed compared to almost 20% in the other two categories, and then finally when it comes to being inactive there aren't that many differences between the marital group, so it is around 9 or 10% depending on group, and as you can see, this interpretation is much more straightforward than the one using odds ratios.

So how do we know which variables we should be, we should include in our model. Well as usual, first when you're formulating your research questions and looking at your data set, you should let theory and relevant background knowledge just as other empirical studies on similar subjects, to inform you when it comes to even which variables you should consider in your model, and then once you've done that and decided a list of variables that you will include you can then go ahead and run your models, and then you can use likelihood ratio tests in the way that I showed you earlier to decide whether that variable, as a whole, is significant in your model, and if it is then you can retain it and when you're reporting your results you may want to comment on the wall tests that show you the pairwise statistical significance for that given category, in that given equation.

To sum up, multinomial regression is something that is a model that you can use for categorical responses with three or more categories, and the categories do not have to be ordered, they can be, but they don't have to be. Interpretation can be conducted with either using log-odds, odds ratios or predicted probabilities like for any other logistic regression model that you might conduct, and as always log-odds are not very intuitive so we don't usually use them. You can use odds ratios but they can get a bit tricky especially for dummy variables because you end up doing these multiple comparisons that I've also shown you today, so it can be easier to understand predictive probabilities but then you need to decide which values you will hold constant for the other variables when you're making these calculations.

And that's all, thank you so much.