Multinominal logistic regression, Part 1: Introduction

Author: Dr Heini Väisänen

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

This video is part of an NCRM Online Resource

Hi, my name is Heini Väisänen and I'm a lecturer in Social Statistics and Demography at the University of Southampton. Today I will talk to you about multinomial logistic regression which is a regression model that you can use when you have categorical response variables. This first part is an introduction to the method just looking at when you might use the method how it works and then giving you a simple example of how to interpret results from the model.

So the outline of the session is as follows, we will first start by talking about categorical response variables so what kind of outcome variables you might use with this type of regression models, and then we will move on to talking about how the model actually works, and then we will look at an example where we study economic activity and its relationship with gender using these methods. And if you haven't yet watched Binary logistic regression videos I would recommend you watch them first before you move on to these sessions.

So what do we mean by categorical response variables. Often we're interested in variables that have more than two categories so three or higher and they are not ordered so the categories that we use we couldn't say which category is larger than the other but we can clearly say to which category all of our respondents or all of our cases in our data set belong to. So examples of these type of outcomes are for instance voting intention so we might have data about what party respondents are planning to vote in the next election whether it's conservative, labour, liberal democrats or some other party in this case we would have four categories it could be if you're a demographer like me and interested in delightful subjects such cause of death so we might be interested in figuring out which characteristics are associated with different types of cancer as cause of death for instance.

The idea behind multinomial regression models is that it's basically an extension of the binary logistic regression model and that's why I recommend that you watch those videos first before you move on to this one.

So instead of just running one model like we do in a binary logistic regression model we model as many pairs of categories response categories as we need. If our outcome variable has R number of categories then there are R times R minus 1 possible pairwise models that we could run. But luckily we don't have to run as many models actually we only need R minus 1 of these models. So if we have three categories in our outcome so for instance if we were interested in whether someone is going to vote for conservatives, labor or some other party we would have six different possibilities of pairing them up so we could pay categories one and two one and three two and three or two and one three and one and three and two. But actually one and two is the same pair as two and one etc so half of these drop out already and actually we will end up only modeling two pairs because we will select one of the categories as a baseline or reference category and only compare the other odd two outcomes to that category and I will walk you through what that means.

So the interpretation of a multinomial regression model is based on this idea that you have a baseline category or it's also called reference category of the outcome. So if you've used regression models before you might be familiar with using dummy variables as explanatory variables and if you

do that you choose one category as a reference category and then you compare the other categories to that reference. And it's basically the same idea here except that you're doing that for the outcome you might as well do it for your explanatory variables but let's not talk about that yet. So the interpretation of this model depends basically on which response corresponds to the numerator and which does correspond to the denominator. The category corresponding to the denominator is called the reference or the baseline category and we compare the numerator to that category. Mathematically speaking it doesn't really matter which baseline category you choose but it usually is easy there are different choices that you can make which might make your life easier when you're trying to interpret the results of the model. So in case your outcome category is ordered you might choose the first or the last category. I know I said earlier that normally we would use nominal variables outcomes here but sometimes we might have ordered variables and we might use multinomial regression to model those. It could be the most meaningful category in your outcome depending on your research question, it could be the most frequent frequent category so you just look at which category has the most responses and then you compare everybody else to that and the only rule that I have against choosing a category is that it's usually not a good idea to choose a rare category as the reference category because that can lead to inflated ratios and difficulties in interpretation of the model. Once you have chosen your baseline category you can then run your models. So staying with our example of a model where we have three categories of the outcome you could for instance choose one of them as the reference category and you would then end up with two model equations. So these are like two binary logistic regression models but they are run at the same time and you compare two of the categories to the third one which is the reference category. So looking at these equations here on the slide which outcome category one two three do you think is the reference category here.

Remember earlier that I talked about numerators and denominators and their relationship to the reference category and as you can see the denominator in both of these equations is category three. So that means that in whatever our outcome is here the one that we coded as outcome number three is the reference category and then we compare categories one and two to that in our models. Next I will move on to looking at an example which makes it a bit easier to understand what exactly is going on here. So our example looks at the association between gender and economic activity our outcome variable has three categories so whether someone is economically inactive so maybe they're on parental leave or maybe they they're retired whether they are unemployed and looking for work or whether they are currently in employment. And then we're looking at how whether you're a man or a woman is associated with your likelihood to be in these different categories.

Here's our data these are real data coming from ONS opinion survey while being module collected in April 2011 and the data source is shown here on the slide and in using these data our outcome variable is distributed so that the most common category is in employment so about 56 of our respondents were working at the time of the survey the second most frequent category was economically inactive so around 39 of our respondents were not working and not looking for work so maybe they were retired on parental leave or students or something like that. And about just under five percent of the respondents were unemployed at the time of the survey. Overall we have 1124 respondents here in our data set. Out of these respondents around 54 were women and around 46 were men.

o when we put economic activity as the outcome in a multinomial regression model we will end up with two equations as shown here and as seen on the previous slide. So one is comparing category one so economically inactive to category three in employment and the second equation is comparing category two which is unemployed to category 3 which is in employment. And we have

chosen here category 3 in employment as the baseline category because it was the most frequent category in our dataset looking at the equations ß1 in the first equation is the effect of our explanatory variable so gender on the log gods of being in category 1 instead of category 3 so being economically inactive rather than in employment. And ß2 is the effect of tender on the logos of being in category 2 so that is unemployed instead of category 3 which is in employment. And like in binary logistic repression we can obtain odd ratios by calculating the exponentiated values of ß1 and ß2 and then use that to interpret our results.

I've used Stata here to run these models you might use some other software but the results are likely to look quite similar to here. So here the results are shown as on the logic scale so you might remember from binary logistic regression that this is a scale that we don't often use to interpret the results because it's not very intuitive but it shows you already some information about what the relationships are going to look like. So if we look at the first set of results here our first equation which is comparing being unemployed to being in employment we have one explanatory variable which is gender man is the reference category and woman is is the category that we see here and we see that there is a negative effect for being woman and being unemployed so that means that women are less likely to be unemployed rather than in employment when we compare women to men. So as you can see these models become quite complicated quite soon because you are comparing so many different sets of variables. So you have the reference category of the outcome variable and since we have a dummy variable as our explanatory variable here we are also comparing men and women. The second equation shows that this time the coefficient is positive so it means that women are more likely than men to be economically inactive rather than in employment.

Usually if we want to say something a bit more meaningful about our results we might want to look at the results in the odds scale. And this is how these results look like in Stata, so it's exactly the same model it's just that I've asked I've asked Stata to show the results on the odd scale rather than ask log odds. And here you can see that for women we have an alt ratio which is below one which means that there is a negative association like we already saw from the log odds and for the second equation um the alt ratios are larger than one which means that there is a positive association.

If we wanted to put these results into English language we might want to say something like this. If we were to interpret these results using the odd scale we could see we could say that the odds of being economically inactive rather than in employment are 73 higher for women than for men and below you can see how I came up with this number 73%. And for the second equation we could say that the odds of being economically inactive rather than in employment are 42% lower for women than for men.

Sometimes instead of using odds to interpret your results you might want to use predicted probabilities most statistical software will calculate these for you but I'm going to show how to do this by hand so that you know where these predictive probabilities are coming from. So if you remember how predicted probabilities are calculated for binary logistic regression you'll see that this is quite similar the only difference is that since we are dealing with two pairs of equations we also need to calculate a higher number of probabilities using more equations than we do in the binary logistic regression model. So basically what we do if we want to know what is the probability of someone of being in category one so in our case this would be an economically inactive we would take the first equations of the equation corresponding to the category economically inactive exponentiate that divide that by 1 plus the exponentiated value of the first equation plus the exponentiated value of the second equation which in our case is the equation for unemployment. If we want to calculate the predicted probability for category 2 it's otherwise the same thing so the

denominator is exactly the same as before but the numerator changes so now in the numerator we plug in values from the second equation which is the one for unemployment. For the reference category the numerator is just one that's always the case for the reference category and the denominator is still the same as it was for the other two categories.

And if you had more than three categories in your outcome then your equations will be even longer than they are now because in the denominator you would always have to have all of your equations shown.

I've shown here how you can calculate the probabilities for women and I've plucked in the values from the table that I showed you earlier the one that showed the results in the log odd scale and we would always use the log odd scale here and then exponentiate that so the first equation calculates the probability of being economically inactive if you're a woman, the second one calculates the probability of being unemployed if you're a woman and the third one calculates the probability of being in employment if you're a woman. And the difference between calculating the probabilities for men and for women is that for women we need to take into account $\beta 1$ and $\beta 2$ so the coefficients attached to the gender explanatory variable which in the first equation is 0.550 and in the second equation is negative 0.537. For men we can drop this because they are the reference category in this dummy variable so their coefficients would be multiplied by zero and when you multiply something by zero it disappears so for men we would only need to exponentiate the constants from these two equations.

And if we calculate these probabilities for men as well we could then summarize our results for instance using these two sentences. Among women the probability of being economically inactive was 46% in employment 51% and unemployed 3%. Among men the probability of being economically inactive was 32% in employment 61% and unemployed 7%. For multinomial regression models predictive probabilities are usually quite useful to use for interpretation because like like you saw with the odds example it can get quite complicated with the multiple comparisons otherwise.

Thank you.