Ordinal regression_Part 1: Introduction

Author: Dr Heini Väisänen

Transcript of: https://www.youtube.com/watch?v=rPcMcW25PPA&feature=youtu.be

This video is part of an NCRM Online Resource

Hi my name is Heini Vaisanen and today I will talk to you about Ordinal Regression. This first part is an introduction to the method and it will tell you when to use the method and what kind of variables we need if we want to use this.

So the outline of the session is as follows, first we will talk about ordinal response variables, which are the type of outcome variables that you would use with these models, then we will talk about what we mean by cumulative odds and probabilities, which are important for these type of models, and then finally, we will look at the model itself, so how does an ordinal regression model work and sometimes this model is also called proportional audit model or cumulative logit model.

So many categorical variables that we use in our research have a natural ordering, so for instance, we might be interested in how severe someone's symptoms are if they're suffering from a disease and we might categorize them into low, medium and high. While it would be difficult to say if there is a distance between these categories, or if the distance is the same between any two categories, we can still easily say that there is an order, low is less serious than medium and medium is less serious than high. We could also examine someone's attitude towards a social question and we might ask people whether they are in favour of something, whether they are indifferent and against and again we have a natural ordering here, going from more agreeable towards less agreeable. If we use this information about how the categories are ordered we can conduct more informative and powerful analysis than if we don't, because we could use multinomial model to model these type of variables, but actually if we use an ordinal model we get a more informative and powerful analysis.

An ordinal model essentially wants to figure out what is the cumulative probability of being in different combinations of the categories, of the outcome, and here's some notation. So when you see Pk, that is the probability of being in the response category K, so I sometimes also call it a response probability. Cpk means the cumulative probability of being in category K or lower, so if we had Cp2 for instance, that would be the cumulative probability of being either in category 1 or 2 combined. And 1-Cpk is probability of being above category K which is basically the inverse of the situation that I just described. If we know all the response probabilities of a certain variable then we can calculate the different cumulative probabilities for that variable, so Cp2 for instance would be P1 + P2, if we know the response probabilities we can figure out the cumulative probabilities P1, so the probability of being the lowest category is the same as Cp1, so the cumulative probability of being in the lowest category, whereas for the other categories we would need to combine the category in question and everything below that, and for the highest response category, so if we had three categories, P3 would be 1-Cpk-1 so Cp2.

Here is an example of how this actually works with some numbers. So let's pretend that we have an outcome variable with three categories and we have some probabilities of each of our respondents being one of these categories. Now the probability of being the lowest category P1 is 0.5, the probability of being the middle category is 0.3, and the probability of being the highest category is 0.2. If we want to figure out the different cumulative probabilities, we can do that now that we know the response probabilities. So the probability of being in the lowest category is Cp1 is equal to just the probability of being in that category, so that is 0.5, so we don't actually need to calculate anything here, we just know that it's 0.5. The probability of being in category 1 or 2, so in other words Cp2, is the probability of being in category 1 plus the probability of being in category 2, so that is 0.5 + 0.3 = 0.8. The category, the probability of being in category 3 or lower, so Cp3, is 1 because 3 is the highest category that we have in our outcome variable. If you don't believe me you can calculate 0.5 + 0.3 + 0.2 and you'll get 1, but we don't even have to do that because we know that everyone has to be in one category so the probability of being in the highest category or lower is always 1 or 100%.

Like with other probabilities we can transform these probabilities into odds and log odds. So if we take the cumulative probability of being category K or below Cpk and divide that by the inverse of the probability, we get cumulative odds, and if we take a logarithm, a natural logarithm of these odds then we get cumulative logits, just like in other regression models, other logistic question models. We can then use this cumulative logit transformation in our ordinal model and on the left hand side you can see that we have a logarithm of the cumulative odds and then on the right hand side we have something that looks quite similar, even if not exactly similar to other regression models that you've seen before. So we have an intercept and we have a slope, so here we only have one explanative variable in our model. So like in a multinomial model, in ordinal model we are estimating K-1 equation simultaneously, so we if we have three outcome categories then we will estimate two different models. The difference to multinomial models however, is that while each equation has a different intercept, Alpha K, the slope in every equation is exactly the same, in a multinomial model the slope would be different in every equation. The intercepts that we get are always ordered in size, which is also different from a multinomial model. So the intercept in the first equation should always be lower than the intercept second equation etc.

You might have noticed that if you look at the regression equation on the right hand side you have something a bit strange going on before the Beta. So you have a negative sign there before Beta1 and that is there to make interpretation of these models easier and this is what most statistical software do Stata and SPSS for instance. It means that once we take this negative sign and plug it into our equation, a positive slope implies that higher values of the exponential variable are associated with an increase in odds of being in a high response category, rather than a low response category. If we didn't have the negative sign it would be the other way around which would be quite unintuitive.

The same odds ratio applies to all of the thresholds, so in all equations for all the different intercepts, because we have made something that is called the proportional odds assumption. So we assume that it actually makes sense to have the same slope for every equation and there are ways to test whether this is actually the case but we will talk about that in later sessions.

Now we will look at a simple example of an ordinal model. We have some data about undergraduate plans to apply for postgraduate stud. So a survey asked undergraduates how likely they were to apply for a postgraduate study and they had three different options that they could choose from, they could say unlikely, somewhat likely and very likely. We also collected data from their parents education and we want to know whether their parents education is associated with their own likelihood of applying for postgraduate study. So the parents education is binary variable which takes value one, if at least one of the parents is a graduate themselves and takes a value zero if they are not. In the table you can see how the outcome variable of the likelihood of, about the likelihood of applying for postgraduate study is distributed, so we have 55% of undergraduates in the unlikely category, 35% in somewhat likely category, 10% in the very likely category and in total we have 400 respondents.

If you use Stata, here is how your ordinal regression results will look like, so these results are in the logit scale for now. In the first row where you have some numbers, you see your coefficients. So the 1.12 here corresponds to the likelihood of applying for postgraduate studies, if at least one of your parents has graduated. Then the next two rows tell you what the intercepts are for the two equations that we have run in this model because we have three categories in the response . So the first of those two rows says 0.376 and that is the intercept for our first equation and then the second row says 2.452 which is the intercept for our second equation.

Like with any logistic model you can interpret these results using odds and like with any logistic regression model you do that by exponentiating your logit value, so if you exponentiate 1.167, which was our coefficient for parent education, you get 3.09, and this 3.09 remember, does apply for all of the two equations that we modelled here. So we could say that for students who have educated parents, the odds of being in the very likely versus the combined somewhat and unlikely categories, are three times greater than those whose parents are not educated.

We could also say that the odds of being in the combined very and somewhat likely categories versus unlikely, are three times higher for those with educated parents compared to those whose parents are not educated. So to sum up, that means that the also being in higher, rather than a lower category of the outcome are higher for students with educated parents.

Like with other logistic regression models you can also calculate predicted probabilities. If you want to figure out what is the probability of being in one of the exact response categories, out of the three possible categories and you can see the formula for that calculation here, and you can see that it looks very similar to other logistic models, the difference, the main difference being that on the left hand side you see that we are calculating cumulative probabilities rather than response probabilities, and on the right hand side you can see that we must remember to change the sign of our slope, so we have a negative sign before the slope. So if you wanted to calculate the probability of being in category 1 for those whose parents are not educated, we would take the first equation which had the lower intercept 0.377, we would take our slope for parent education 1.127, we would change the sign because of the negative sign here in the equation, and then we would multiply that by zero because now we are looking at the reference category of this exponential variable, so those whose parents were not educates. If you were looking at those whose parents were educated you would multiply it by 1. We exponentiate that equation, divide that by 1+ the same thing, so exponentiated value of the first equation. If we solve that, we get 0.59, and that 0.59 is both the probability of being in the exact category 1 and the cumulative probability of being in category 1. For category 2, we do the same thing, except that we change the intercept now for the high intercept that we had in the second category and then when we solve that equation we get 0.92, so now we know that the cumulative probability of being either in category 1 or 2 is 0.92, but we don't yet know what is the probability of being in the exact category 2. However, we can fairly easily find that out by taking the cumulative probability 2 that we just calculated and then subtracting the cumulative probability 1, which we also just calculated, so we plug in the values 0.92 - 0.59 = 0.33. So now we know that the probability of being second category for those whose parents were not educated is 0.33. We can also calculate the probability of being the highest category using the information that we already have, so the probability of being in category 3 is 1 - the cumulative probability 2, 1 - 0.92 is 0.08.

Here is the table of response probabilities showing the response, the different probabilities for the two categories, so those whose parents were not educated, so the ones that we just calculated and those whose parents were educated. I haven't shown how to calculate that, but the only difference is that you multiply the slope by 1 rather than pie 0, and you can see that parents education does make a difference, so those whose parents were educated had a higher probability of being very likely to apply for postgraduate education, 25% versus 8% among those whose parents were not educated, for instance, and if you look at the lowest category, unlikely to apply for postgraduate studies, we can see that for those whose parents were educated and the probability of being in that category was 32% compared to 59% among those whose parents were not educated. When you use ordinal models, calculating response probabilities is usually quite useful because sometimes dealing with the cumulative odds is a bit difficult to understand if you're not very familiar with this model, so depending on your audience you might want to consider using probabilities rather than odds when when making interpretations from these models. Thank you!