Binary logistic regression: Multivariate binary logistic regression (Video 2 of 3)

Hi everyone, my name is Heini Väisänen.

I work at the University of Southampton and today I will talk to you about multivariate binary logistic regression.

So the outline of today's lecture is here. So we will talk about what we mean by multiple logistic questions we will talk about model selection so how to decide which variables to include in your model and how Wald tests and likelihood ratio tests relate to model selection.

So when multiple logistic regression we mean that we mean binary logistic regression models where we have more than one explanatory variable like in other regression models you can add more than one variable and that indeed is usually the interesting part about conducting regressions rather than crosstables for instance. The interpretation of the variables is similar to what you saw in the first video where we had just one model, one explanatory variable in our model, but we are conducting this invitation while we are controlling for or holding constant the other coefficients.

So normally we would try and add variables that are meaningful or might impact the outcome that we're interested in, but which are not always the main part of interest with respect to our research question.

Like with binary logistic regression models where you only have one explanatory variable interpretation can be made in three different scales. Log-odds scale, odds-scale and predicted probability scale. And we will talk about the odds-scale and the predicted probability scale today but we will not talk about the log-odds because as we discussed in the first lecture it is not a very intuitive way to interpret this regression.

So, I will approach this subject with an example. So I extracted some data from the demographic and health survey of Ghana from 2008 and I was interested in studying which characteristics of women or mothers are associated with the likelihood of having had received, having received assistance from a healthcare professional in their most recent birth and I have the following characteristics of rolling variables available in my data set so we know the quintile of wealth of the household where the woman lives going from the poorest 20% up to the richest 20%. Then we have, we know the number of children that she has, we know whether she lives in an urban or in a rural area and then we know what level of education, what is the highest level of education that she has, whether that's no education at all, primary, secondary or higher education.

Here is an information about how these variables are distributed in our data set. So when it comes to our outcome or our dependent variable, so whether the woman received assistance from a healthcare professional in their most recent birth we have 58% of the women who received assistance from a healthcare professional and um the other side of the coin is that 42% did not receive a help from a healthcare professional in their most recent birth. When it comes to the explanatory variables their mean number of children in sample was 3.5 and we have 2144 women in our sample so just keep in mind that this is not the mean number of children inthe population of Ghana in 2008 because by definition we only have women who have at least one child in our analytical sample because we are interested in their experiences at their most, in their most recent births. So, you have to have one child by definition to be included in our sample.Around 35.6% of the women lived in an urban area and the other side of that coin is that around 65% of women lived in a rural area.

When it comes to education around 36% of women didn't have any education. Around 24% had primary education, around 38% had secondary education and around 2% had higher education. When it comes to wealth in our analytical sample, the poorer health categories are more populous than the richer categories which might be linked to the fact that we are only looking at a selected group of people so those who have at least one child. When we put all of these variables into a regression model in the binary logistic regression model. Here are the results. So we have first the urban rural residence, then we have wealth, then we have education and finally, we have number of children as a continuous variable.

All the other explanatory variables are categorical. When it comes to the columns that you see in this table the first column that shows you results, shows you the results in the log-odds scale so the unintuitive scale but the one that is quite similar to linear regression. Then the middle column shows you odds-ratios and the last column shows you P-values from wald tests and I will come back to what that means. Usually regardless of which software you use, whether you use data or R or SPSS or some other statistical software you will get lots of other things as well in your model like confidence intervals and standard errors and those are important as well but today we will only talk about these three, the log-odds, odds ratios and p-values so that's why I'm not showing the other R values.

So how do we interpret our model? what do these numbers actually mean? We could start by looking at the odds-ratios and the interpretation in our scale. I'll give you a couple of examples that give you different types of variables so I give you one categorical and one continuous variable and also different directions of effects so positive and negative effects. So let's look at wealth first. If we wanted to see, focus on the differences between the poorest category and the richest category we could say that when controlling for education, number of children, and place of residents those in the richest quintile have 13.7 times the odds of having had assistance at the most recent birth than those in the poorest quintile.

And how do we know this? Well if you look at the odds-ratio column in this table and you look at wealth richest row you can see that the value there is 13.75 so it's just directly taken from that column and since the poorest is the reference category it means that when we compare the richest and the poorest which is almost 14 times as likely to have had assistance in their most recent birth while we are here we might observe that the odds ratio for number of children is 0.95 and if we look at the corresponding log-odds value we can see that it's a negative number so we know that when the number of children goes up then the odds of having had assistance at birth goes down. And if we want to express in odds radios or in odds scale how much exactly does do the odds reduce with each additional child we could say that when we are controlling for the other variables in our model each additional child reduced the odds of having had assistance at the most recent birth by 5%.and the 5% comes from the calculation that I showed you last time that you can use when you're doing interpretation in odd scale so you take the odd ratio or 0.95 you subtract 1 from that so 0.95 minus 1 and then to turn that into a percentage you multiply that by a hundred and we get negative five so that means that for each additional child you ought to reduce by 5%. More generally when it comes to the interpretation in odds ratios what we do when we interpret these model results in the odds scale, we calculate the ratio of the odds of the categories of interest. So the ratio of the odds of the poorest category and the richest category for instance for the wealth example that we just saw. And that's why they are called odds ratios or ORs. For categorical variables we calculate the odds ratio so that we compare the other categories to one reference category and this is something that you already know from logistic regression it says that the difference, sorry from linear regression it's just

that the difference multiplies rather then it's additional like in linear regression for continuous variables, the odds ratio expresses how much the odds of the outcome being equal to one so in our case and the woman having had resistance at birth increases when the continuous variable increases by one unit. And as soon as you have the log-odds scale results you can calculate the odds ratios by exponentiating the estimates on log-odds scale.

Like we discussed last time you can also interpret this results on probability scale if you calculate something that we called fitted or predicted probability. And the formula for doing that is here so we figure out what the probability or Pi is by exponentiating the sorry [cough] the equation of interest dividing that by 1 + the exponentiated value of the same equation as above. So when it comes to the equation of interest we have to decide what we want to calculate so as you might remember the probabilities these fitted probabilities depend and are very different depending on which values of these planetary variables we choose. So we might be, if we have a lot of continuous variables in our model, we might want to put everything in there as there means and kind of calculate the average predictive probability. If we have categorical variables maybe we want to choose the most frequent category to kind of find this average experience or there might be other values of interest that would be relating to our research questions. Or we could, if we're using software to calculate these probabilities we can leave them as observed but that's not something that I will talk about today but if you do the exercises related to this lecture then you will see how that works.

So I will just give you an example of how you would calculate these probabilities by hand. So let's say that we're interested in knowing what is the probability of having had assistance at most recent birth if the woman is living in an urban area, belongs to the poorest wealth category, has no education and has two children. And when we plug in the values in this equation we need to remember to use the log-odds scale. So just to remind you here's the table again and I've highlighted the values that we need. So for a place of residence, the woman lives in an urban area so the log-odds value that we need is 0.95. For wealth poorest category is actually the reference category so it means that its corresponding log-odds value is 0 so that cancels out from our equation.

Then for education, we have -1.55 and for number of children we have -0.06. So when we plug these in to the equation first we start with the constant which in our case is 0.53 then we plug in the value for living in an urban area or 0.95 and then because that is a categorical variable we multiply that by 1. Then we plug in the value for having no education -1.55 and again that is the categorical dummy variable so everything is 0s and 1s so this is 1 we multiply by 1.

And then finally we want to know the probability for someone who has 2 children so the -0.056 for having children we multiply by 2 because there are 2 children then we exponentiate this divided by 1 plus the same equation and if we solve this equation we get 0.458. So that means that the probability of someone who lives in an urban area is in the poorest wealth category has no education and has 2 children and the probability that they had assistance from a health professional in their most recent birth is around 46%.

So it's just to sum up. log-odds, we can do interpretation in log-odds scale but it's not very intuitive. Odds ratios tell us about the relative differences and apply to the entire scale of a continuous variable. Probabilities tell us about the absolute levels of risk and when we calculate them we need to decide at which values the other variables are held

It does not apply to the entire scale of a continuous variable but it is often the most intuitive option so you might want to calculate the range of probabilities rather than just one like I did here and usually you wouldn't do this by hand but using software.

Then a few words about model selection. So I haven't talked about p-values at all yet so let's take a look at what these mean in a logistic fashion context. So the thing that you are normally see printed when you conduct these models using software is something called Wald-test so it's very similar to t-test in our last regression so you kind of already know how to deal with it. And especially for continuous variables and binary categorical variables it is a very useful way to assess significance. For dummy variables, unless you're conducting a joint Wald test it is slightly less useful because it only tells you the significance of an individual category from the reference category. So what do I mean by that?

Here we have the p-values from our model and if we look at the first variable, place of residence, we can see that that is associated with a very small p-value. And that means basically that place of residence is important in small model and there is the difference between urban and rural populations in the population, or it's very likely that this difference also exists in the population.

When it comes to number of children, so the last variable that we added to the model which is the continuous variable we have a p-value of 0.021 which at the 5% level again tells us that number of children is significant in this model which means that we expect this association to also exist in the total population. However if you wanted to look at wealth or education we wouldn't be able to say just by looking at these p-values whether the dummy variable as a whole is significant especially for education where we have a mix of large p-values and smaller p-value. So no education p-value 0.041 but then for secondary education 0.276 so we just, the only thing that we know so far is that no education is significantly different from higher education but secondary education isn't significantly different from higher education but we're not sure if education as a whole is important in our model.

But we could use the likelihood ratio test to figure out whether education or wealth matters and what this how this works is that whenever we conduct a binary logistic regression model we do that, we find the estimates by maximizing the log-likelihood or LL and the higher the log-likelihood is, the better the model fits the data so the better the model is. If we take this value multiply it by -2 we get something that we can use to test differences between two models. So if we calculate the -2LL value for two nested models and by nested models I mean that there are some variables that are in common and then one of the models has more variables than the other this difference follows the Chi-squared distribution which you might already know from having conducted Chi-square tests and cross tables.

As you might remember this distribution is different depending on how many degrees of freedom you have and in a logistic regression context we calculate the degrees of freedom by looking at how many parameters we have in each model and the difference in the number of parameters. The more complicated value one will always have smaller -2LLl value which means that the fit is better but this reduction in this value is not always statistically significant and the likelihood ratio health

test tells us whether it is significant and whether we should keep that more complex model or go with a simple one.

So how does this work? let's take a simple example with model one with what with the same Ghana example in model 1 we just have number of children as an expansion variable and in model 2 we have number of children and both as exponentially available variables. And in the likelihood ratio

test we are testing the following hypotheses the null hypothesis is that there is no difference between the models and the alternative hypothesis is that there is a difference between the model.

Here are the results if you calculate by hand so the likelihood value for model 1, so the one that only has number of children is minus -1410 or 1410 if we multiply that by 2 and change the sin we get 2820 for model 2 the-2LL value is 2285.88 if we calculate the difference between these two we get 534.36. We have 4 degrees of freedom because our wealth variable was a dummy variable with 5 categories 4 of them which actually appear in the model because one of them is the reference category poorest and it doesn't do anything in the model other than being the reference category. And if we compare that to the Chi-square square distribution with 4 degrees of freedom we get a very small p-value so we conclude that adding wealth to our model was statistically significant and we should keep it in the model.

So just to sum up. likelihood ratios is often the most important tool that you will compare models especially if you have interactions or big categorical variables and normally you would do this using software as you will see from practical's associated with this lecture. You can also use the Wald test that is automatically printed for you by most software to get information about statistical significance and other than statistical significance remember to use theory and previous research in guiding you when you're selecting which variables to include in the first place and in which order.