Binary logistic regression: Multivariate binary logistic regression (Video 1 of 3)

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 binary logistic regression which is a model that you can use to study outcome variables that are surprise surprise primary. This is an introduction to what the model does and what kind of variables you can use if you're using these models and how to interpret the results of a simple binary logistic regression analysis.

The outline of today's session is we will first talk about binary responses so what kind of variables we might have as outcome or response variables with these modules, then we will talk about link functions so the trick that we use in order to be able to use regression models with outcomes that are not continuous, then we will talk about what the logistic regression model actually looks like and then we will look at a simple example of such model and use that to learn how to interpret the results. So binary response and regression models how does that work. Often in social sciences and other sciences as well we're interested in relationship between categorical variables. So we might want to study things like whether someone has a certain disease we're interested in or not so it could be whether someone has COVID or not and what are the characteristics that are associated with higher risk of getting the disease, it could be a voting intention so we might want to know what is the likelihood that someone votes for a certain party or some other thing in in the referendum for instance, or it might be some other type of categorical variable like likelihood scale that we might have then recorded into binary variables.

In this session I will talk to you about binary logistic regression models which can be used to study the association between first of all a binary outcome or dependent variable so a variable that only takes two values and explanatory or independent variable of any kind. So the explanatory variables can be continuous they can be binary as well they can be nominal they can be ordinal whatever you have in your data set that is fine the only requirement is that the response or the outcome variable has to be binary so only take two values.

When we have binary variables we normally code them as zeros and ones and the way that we code them doesn't really matter mathematically that much but usually it makes sense to code them so that the type of response that you're interested in is coded as one so it might be success of some sort so whether someone passes an exam or not or it might be yes to a binary question it could be whether you've had COVID19 yes or no and we would code yes as one and no as zero and what we're ultimately interested in when we run these type of models is to figure out what is the probability that Y equals one. So what is the probability that someone passes an exam what is the probability that someone has COVID19 and what characteristics are associated with that probability. And we use the Greek letter phi to denote those probabilities.

So why can't we use linear regression why do we have to have this new regression model type that we're using for binary outcomes. Well if you had data where your outcome variable was binary and you had some exponential variables as well and you put those data into a statistical software like SPSS or Stata and you run a linear regression or or less regression model using those data it would give you results and that type of model would is called a linear probability model. The problem with those models is that we are violating some of these assumptions of all the regression models and we might run into problem with predicted values. So as you might remember from linear regression you usually have a continuous outcome that can take many different values sometimes any values and they are not if you then use your model to calculate fitted values from it they are not bound into any interval. However, when we are dealing with binary variables what we want to model is the probability that that Y equals 1 and that's why if we calculate repeated values we should only get values between 0 and 1 because probabilities can't be smaller than 0 or they can't be larger than 1.

However, linear regression does not necessarily make sure that this happens and you could end up with probabilities that don't make any sense. The other problem is that when we run a linear regression when we have a binary outcome variable is that we violate the assumption of constant variance so in a linear regression model we assume that the errors of the of the model vary constantly but if we have a binary outcome that is not true because the variance will depend on the explanatory variables to their influence on the probability. And when we violate this assumption it means that our standard errors are not valid and then any conclusions we make from them for instance about regarding statistical significance will be misleading.

So what can we do in order to make a regression model work for a binary outcome. The solution is link functioned so we can still keep the stuff that you've seen before in linear regression models on the right hand side of the model where you have a constant and then you have estimates attached to variables that tell you how the variables you've entered in your model are associated with the outcome. The only thing we have to do is we have to transform the outcome which is the binary outcome in this case so that it becomes linear and makes sense in the regression context and that transformation is called a link function.

And what the link function does is basically it makes sure that we get rid of the problems that I just described that we might run into if we use linear probability models so if we use or less regression to model binary outcomes. So the link function makes sure that if we calculate fitted values from our model they will stay between 0 and 1 so there will be probabilities that make sense and they will also make sure that when our linear predictor increases then the probability goes closer to 1 and when it decreases the probability goes closer to 0.

Here's the same thing in a graphic format. So the link function that we use in binary logistic regression model is the logic link function. There are other link functions too but we will not talk about them now. And because we use a logit link function the type of repression model that we use is called logistic regression and basically what the link function does is it transforms the probability that we're interested in so the probability that Y equals 1 into a logit transformation. So we take the probability so let's say the probability that someone has COVID divide that by the inverse of that

probability so that someone does not have COVID then we take a natural logarithm of that calculation. This calculation inside the brackets here is called odds. You might be familiar with odds if you ever done any betting and then when we take a natural logarithm of these odds then we have our logit link function. The fundamental concepts that you should keep in mind is that the probability phi is the thing that we're ultimately interested in but we transform that first into odds so we divide the probability by the inverse of that probability and then when we take a natural logarithm of that those odds then we have our link function.

And when we do that we now have a function that makes sure that if we calculate fitted probabilities or fitted values from our model we won't get probabilities that make no sense. So on the x-axis in this graph you have probabilities and on the y-axis you have a linear prediction so that would be the type of predictor that you know you've seen in linear regression models before and as you can see on the y-axis in our case here the linear predictor varies from five to minus five so it can take values outside and 0s and 0 and 1 but when we look at the probability in the X-axis you can see that it's the ace between 0 and 1 acetylene predictor goes down the probability class gets closer and closer to 0 but it never quite reaches it and when the linear predictor increases then the probability gets closer to 1 but never quite reaches it. And the linear predictor can take any values any negative or any positive values in here it happens to be between -5 and 5 but it could be some other values as well.

So here are some characteristics of these different transformations of the probability that you should keep in mind. So first if probability is 0.5 so if it's equally likely that the event of interest happens or doesn't happen then the corresponding odds are 1 because if you divide 0.5 by 1 minus 0.5 so if you divide 0.5 by itself you get 1 and the corresponding um logit is 0 because if you take a natural logarithm of the of number 1 you get 0 if the probability is higher than 0.5 so if it's more likely that the event of interest happens than that it doesn't then your corresponding odds will be larger than 1 and your corresponding logit will be larger than 0. If the probability is smaller than 0.5 so if it's less likely that the event happens then that it doesn't happen then your odds will be smaller than 1 and your logit will be smaller than 0 so it will get a negative value. So this means that logit can take any values any negative or positive values but the probability is constrained be between 0s and 1s. The odds can take any positive values but they can't take any negative values. The probability can never be exactly 0 or 1 so we might approach 0 or 1 but it can never be exactly that.

So let's look at an example I find that that is usually an easier way to understand what's actually going on rather than just thinking about these things in the kind of very abstract math abstract way. So we have some data about birth weights here that we can use to construct a very simple model a very simple binary logistic russian model. So we want to know what is the probability of a normal birth weight associated with gestational age so the number of weeks that the fetus has been in the wall so our response variable or our outcome variable is birth weight coded 0 if it's low and coded 1 if it's normal because our research question asks what is the probability of normal birth weight so that's why we have coded normal as one. Our exponential variable is a continuous variable gestational age in weeks so it is the number of weeks and that the fetus has been in the wall and we have defined low birth weight according to WHO so if the birth weight of the baby was less than 2.5

kilograms or 5.5 pounds then that is low and if it was higher than 2.5 kilograms or 5.5 pounds then that would be normal birth weight.

Here is distribution of our outcome variable so the distribution of birth weight so we have 17 babies that were born with normal birth weight and seven babies that had low birth weight if we want to express that as a percentage we have about 71% of babies that were born with normal birth weight and about 29% of babies that were born with low birth weight in our data set in total we have 24 observations.

So here are the results of our model so if we put these data into statistical software and then we run a logistic regression model and that is predicting the probability of normal birth weight using the information about how many weeks the baby has been in the womb these are the results. So on on the left hand side left hand side we have the logarithm of the odds of normal birth weight which we have noted here by logit and then the probability the probability is wearing a hat because it is an estimate rather than a value that is coming from the total population of all babies born and then on the left-hand side we have the regression equation -48.9 which is the constant +1.31 which is um the estimate associated with our gestational age so if we wanted to start making sense of this equation what does it actually mean.

There are three different scales that we can use to interpret the model. The first and the easiest one is interpretation on the logit scale so this is the same thing that you do in a linear regression model so you just take the 1.31 which is the estimate associated with gestational age and say that every one week every increase of one week in gestational age is associated with an increase of 1.31 in the log of the odds of normal birth rate the problem with this approach is that it's a bit difficult to understand if this is a lot or not because we are not used to thinking in this scale.

Alternatively you could interpret this on probability scale and if you do this you have to choose which value of the explanatory variable you are interested in and you can compute these probabilities for any value of your explanatory variables so any value of gestational age. So for instance if we wanted to know what is the probability of normal birth weight at gestational age week 39 we would plug in 39 to our model here to our equation here then we would exponentiate that equation divide that by 1 plus again the exponentiated value of the equation and if we solve this we get 0.1. So we could report that the predicted probability of normal birth weight at week 39 is 0.9 or 90%.

Finally you can also interpret your results on odd scale. So that means that you take your estimated value so 1.31 in our case for gestational age and you exponentiate that value and that gives you the odds of the outcome happening. So if you exponentiate 1.31 you get 3.71 and that means that with every one week increase in gestational age the odds of normal birth weight increase 3.7 times. You could also express this as a percentage by calculating 100 times 3.71 which is the odds value minus one and you could then report that with every one week increase in gestational age the odds of normal birth weight increase to 171%.

So to sum up, you have three ways to report your results, you might use the log of scale which is similar to oralist regression but the problem is that that scale is not very intuitive and that's why we don't often use it. Alternatively you could calculate a predictive probability and that is actually quite often used because it's easier to understand but the problem with that is that it is specific to the values of exponential variables you have used to calculate your probability so in our case um we calculated probabilities per week 39 if we had calculated the probability for week 36 we would have had a very different answer of 0.15. Finally you could report your answers as odds and that would tell you how the relative level of risk changes as testation weeks increase and you could report that every one week increase in gestational age increases the odds of normal birth weight by 271 percent.