# Multilevel models to study intersectionality

## Transcript MAIHDA - R example (video 5)

Part of the resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20849>

So, in this practical session, we're going to go through an example that uses simulated data based on an outcome HbA1c, which is a measure that's used to diagnose diabetes. So, really this is a study that's looking at diabetes outcomes across society.

 This is based on a paper that is a tutorial and there's a link to that in this in this document. So, if you want to know more about this and see it in a bit more detail, do go read and read that paper.

 So, the first thing is to load some packages and then we're ready to load the data, so I'm going to run this line, which reads in the data set which is a .dta data file, but it reads that in, and we can now look at that data and we can see we've got our HbA1c, our outcome and then we've got that binarised into whether someone is diabetic or not. We're not going to use that variable. We've then go five different variables which are going to define our intersectional strata. So, we're going to be looking at inequalities between combinations of these five variables.

 The next thing that we're going to do is create our strata level variable. So, this is going to be creating a unique identifier for each combination of those five variables. So, we're going to do that by saying that each variable can be defined based on one number in a kind of five digit number where the first digit represents sets, the second digit represents race, the third education, the fourth income, and the fifth age.

 So, if I run this code, we will see that in this data set we've now got this new data set, new variable, which basically mushes together these five numbers into a single code there.

 The next thing we're going to do is import our value labels. So, at the moment each of these variables, we can see this here, are the variables that have character labels, but are themselves characters. So, what we're going to do is turn these variables into factor labels, factor variables using the unlabelled command. So, if I run that you can see these are now appearing with labels that kind of makes sense, so we can see that stratum 21114, the 2 represents female, 1 represents white, 1 represents less than high school, 1 represents low income and 4 represents age 60 plus.

 We might, at this point, want to do a little bit more investigating of this data, so it might be worth making a table of our strata so that we can see from that some of our strata have quite small numbers, even some that have only one observation. Others strata have a lot more into the quite high hundreds. And that's pretty normal for social data where often some of these combinations of these five variables are relatively rare.

 We could add that to our data set as a variable if we wanted to. If I run this code, we're saying we want to group by stratum and then create a new variable called strataN which is the number of observations in each stratum. So, that's now appeared there on the right hand side.

 We could also look at our independent variables, our variables that define our strata, get a sense of the number of observations in each, check these make sense, so check that we've got kind of similarish numbers of men and women, check that these proportions look how we would expect for the population that we're looking at, check that out factors are what we're expecting. So, this is based on American data where the division between white, black and Hispanic probably makes the most sense. We've got four categories of high school education, four categories of income, and four categories of age. We probably also might want to get some summary statistics of HbA1c, check that these look kind of how we might expect in our data and we might also want to plot that outcome variable in histogram. So, what we've done here is use ggplot so that we want to use data where the variable we're looking at is HbA1c, we want to make a histogram, we've converted that X variable here to a percentage, and then we've got a few kind of aesthetic things that we've added to this graph, such as the maximum, the minimum, the sub, I mean, etc.

 Okay, we're all ready to start our modelling and we're going to start with our initial model, which is our null model. So, this is the model that contains no fixed part X variables. We've just got our outcome and our random effects. So, we're going to use the command lmer, which is either the LME4 package, that stands for linear mixed effects regression. We then specify our outcome variable, HbA1c, and then on the right hand side of our equation we've just got our strata level random effects. So, we're saying we want these random effect defined by strata and the thing that we want to vary is the intersect represented by one there. And we’re saying what data we're using and running that model. We’ll look at the output for that in a second once we've run our second model as well.

 We can make predictions based on this model, so I'm just using the predict command that will make that prediction for each of these, each of our strata, based on the fixed part of our model, which at the moment doesn't have much in it, and the random level 2 part of our model.

 Next we can run our second model, so this is the model that's the same as our first model, but we've now added in the additive effects of our strata variables, okay, so it's the same code as Model 1A, we've added in code for sex, race, added in variables into the fixed part of our model there. These are already factor variables, so R will automatically know to treat these as factor variables, as categorical variables, it will remove one, our dummy variables for each category minus a reference category.

 Done that. So, we can now make some predictions. I've already done this. This takes a minute or two, so I'm not going to run it again. We're using the predict interval command and this is a really useful command because not only does it give us predictions as the predict command does, but it also gives us confidence intervals around that prediction. So, that's a really useful kind of feature of this.

 And that produces a separate data frame where for each observation we've got a different prediction and a different upper and lower bound on that prediction along with an ID there. Actually, the ID we have on later, so the ID we add into this data frame here, it’s on the mutate function here. This is going to allow us to merge this into our original data. So, I’m going to run this to add that ID variable, which looks like I've already done. I’m going to do the same with our tut data set, we add that ID variable, and then when we merge these two together using the merge format, we can just say merge the original tutorial data set with those predictions by this newly created ID variable.

 And then we’re just going to rename some variables just to make it clear which model these are based on. So, we're just renaming that variable called Fit, m1Bmfit, and the same for the competence intervals.

 The next thing we're going to do is create a new data frame which has data just at the stratum level. So, at the moment we have this big data frame which has an observations for each individual, but we've now we've got these predictions, we don't actually need that level of detail because actually all the predictions for anyone in a single stratum are going to be the same. So, we're going to collapse this data down to the strata level. So, we're going to create a new data frame called stratum level, and to do that we’re going to aggregate our variables. So, only variables that change in our data set, we need to include here. At the moment, that's just that we're interested in that just HbA1c, and we’re going to do that by any variables that don't change within stratum which is pretty much all of our other variables here.

 Okay, let's have a look at our model outputs. So, I'm going to use tab model to display these because it presents them quite nicely in a table. And we can see that here. So, you can, bigger. There we go. So, we can see here we have our, in the first model, an intercept of 40.79. That's effectively our mean value of HbA1c across all stratum. In that we've got a level 2 variance of 9.37, that's the variance between strata, and a level 1 variance of 90.26. That's the variance within strata between individuals. So, we can see about 10% or about 9% of this variance occurring at the higher level. So, that's our estimate of what's labelled here is the ICC intraclass correlation, but it's also the variance partitioning coefficient, the BPC. Those two things are equivalent.

 And then we add in our main events. We can see these are kind of statistically significant in the directions we’d probably expect, so women have a lower at risk of diabetes than men, black and Hispanic people have a higher risk of diabetes than white people, more educated people have a lower risk of diabetes, high income people have a lower risk of diabetes and older people have a higher risk of diabetes.

 We can see that by adding in these main effects, our level 2 variance has been substantially reduced. Okay, it's gone from 9.37 down to 0.88.

 So, that's a reduction of the level 2 variance of about 90%. So, our PCV, our proportion change with the variance, is about 90%, yeah. So, most of our variance is explained by those additive main effects, but there's still a substantial chunk that's unexplained, that 10%, which is a result of multiplicative combinations of these various. 90% is about standard I would say for most MAIHDA analyses in health, that's the kind of area that we get. So, if you're finding PCVs that are lower than that of 80%, 70%, that suggests quite a lot of multiplicativity. If you're finding PCVs of 95%, 98%, that's perhaps suggesting there's not so much multiplicative going on as an additive stuff really is the kind of the dominant in what produces these intersectional differences.

 So, now it might make sense to make a plot of what our model is predicting for our residuals, sorry, for our strata. So, to do that, what we're doing is we're taking our strata level data set, we're going to create a variable which will give us a rank of our predictions.

 So, if I run that, I just have a look at that stratum data, you can see we've got our Model 1B predictions. Those are our predictions for Model 2. And then we've got out the rank of those predictions on the right hand side there as well.

 So, we can then make a plot of those predictions against that rank. We're going to make it do that as a point graph and then provide a range around those points that give us our confidence intervals. So, if I run this code, that gives the graph here, which is perhaps quite difficult to interpret on its own. It kind of gives us a sense of which strata, there are some strata that are predicting quite low, other strata that are predicting quite high. It gives us a sense of the range, but it's quite difficult at this point to kind of say what each of those strata are. This graph doesn't have labels for each of those points. It might be sensible to look at which are the highest and lowest stratum, so the highest, most predicted stratum, so we’re going to sort by rank and then look at the top and bottom. In this case, the top and bottom six stratum. And from this a few things are kind of popping out, I suppose. You can see that the majority of those top ranked stratum, the lowest at risk, these are the people down here, they're mostly white, they're mostly female, they're mostly college educated to some extent, they're mostly high income and they're mostly young. Okay, so, we can see that the most advantaged groups tend to have those characteristics, but the least advantaged groups are quite mixed in terms of those sets, but they're all black, none of them are fully college educated, they're all old or older than 60. So, we've kind of seen by looking at these the kind of variables that are perhaps really dominating in terms of what's going on with the strata.

 It might be that we're also interested in just looking at the multiplicative effects, okay, so these are the effects on top of those additive effects. So, we know that there are additive effects that we saw in our output here. We saw that there were an effective set, an effective ethnicity and so on. And those additive effects would suggest something about the predicted values of HbA1c. But we also know that there is something additional on top of that as well. So, we might be interested in particular in plotting those individual, those multiplicative effects. And we could do that using the REsim function which extracts the random effects from our model and then the plot REsim will automatically plot those as well. So, if I run this, I've already run the first line, so if I now run this plot.

 Yeah? There it is. So, you can see in this plot there are a few strata which seem to be less disadvantaged, more advantaged than we would expect given their combination of additive effects and a few strata that are more disadvantaged given the additive effects.

 This again is a graph that’s quite difficult to interpret. We can see that low strata can form approximately to that additive effect, but the ones that are highlighted are coming out of significantly statistically significantly different from those. But we probably want to identify which ones those are, so we're going to now make a plot that just highlights those that are statistically significant. So, we're going to use what comes out of this plotting process. So, when we do this, within this plot object P, we have a lot of data and we can extract that data and included in that data is data about statistical significance. So, we can first extract that data, we can then filter it so that only the statistically significant strata are included, filtering based on significance being true. I've done that.

 And then we can plot just those strata there, just those strata that have statistically significant multiplicative effects. And I've added, in doing this, I've added in the labels of those strata as well. So, we can see that strata 22114 is particularly advantaged. Okay, that strata seems to be particularly advantaged in terms of its risk of diabetes at the generally lower level of HbA1c. A bit bigger. And this strata 21223, we can see that that's particularly disadvantaged. That's the strata that has higher levels of HbA1c than we would expect.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK