# Multilevel models to study intersectionality

## Transcript MAIHDA - Modelling options (video 2)

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

Welcome to part 2, modelling options and the MAIHDA approach. I’m George Leckie from the University of Bristol. So, we’re going to walk you through the actual statistical models underlying the MAIHDA approach here and we’re going to contrast that with some simpler regression, linear regression-based approaches. So, let’s get going.

 So, remember what we’re really interested in is how individuals’ outcomes vary across intersections, where these intersections are combined by combinations of individuals’ social identities, maybe gender, maybe ethnicity, maybe social class and so on. So, the simplest kind of analysis that people would do in that situation would be to run a linear regression model, so we’re going to go through three approaches or options here. So, the simplest thing you might do is the equation given here, so suppose the individual outcome of interest was individual health, it might be something like body mass index or something else, we do a linear regression of that on the social identities of interest, maybe gender in this example, maybe ethnicity, maybe age and so on. And you’d be able to estimate the main effect of each social identity having adjusted for the others, and that would be one way to start to study differences in the outcome across social identities.

 So, that will tell you the effect of each social identity in isolation, but it’s not really directly speaking to how combinations of these social identities lead to a kind of complex mapping of individual health outcomes, so we’re losing some nuance here. This is a bit blunt. So, to make that clearer, let’s go back to the graph that Andy showed in the previous presentation. Essentially, what we’re doing here with a traditional linear regression with main effects is the analysis on the left, and I might even get my little drawing crayons out here. So, we’ve implicitly done this kind of analysis. Okay? But what we’re saying is that often the truth might be more nuanced and more complex. And particularly when you are trying to speak to particular intersections of social identity, so you want to make statements about the experiences of, for example, black women versus white women, and black men versus black women or versus white men, when you want to do that nuanced description across intersections, you really need to at least explore whether the patterns which are really out there in the real world are more complex than what a naïve main effects analysis would suggest. And of course in this particular example that Andy talked through at length, there were distinct differences. Really it was only black women who had distinctly different outcomes from the other three groups, which is the plot on the right, whereas the simpler analysis would be misleading because it would suggest that white women have different experiences from white men and that black men have different experiences from white men, and that was not the case, and a more nuanced analysis would allow for that.

 So, how do we do a more nuanced analysis? That’s where we’re going to next. So, let’s get rid of my children’s scribbles, shift onto the next slide. Okay, so this is what a conventional person doing linear regression would do if they wanted to allow for more nuance and move from the plot on the left to the plot on the right in the previous slide, they’ve put in the interactions between the social identities. Now, in our previous equation we had three social identities. Let’s go wild with the colours, let’s choose a lovely orange colour in this slide. So, my three social identities were gender, ethnicity and age, so with three variables you need to put in the different interactions. You’ve got the two-way interaction between gender and ethnicity, you’ve got two-way interaction between gender and age, and then you’ve got the two-way interaction between ethnicity and age, and finally you’ve got the three-way interaction between gender, ethnicity and age. And you can see hopefully already this is becoming quite a cumbersome model. Okay? Imagine if we added a fourth social identity, there would be further two-way interactions, three-way interactions, and a four-way interaction would be introduced and so on. So, very quickly this approach of putting in all possible interactions becomes very hard for the researcher to interpret and very hard for the researcher, once they’ve worked out what was going on, to present the results in an accessible way to an audience and so on. And in more complex models, maybe you’ve got a count outcome or a binary outcome, some of these cells are pretty small , it all becomes a little bit unstable and noisy and so on. So, really there are limits to the conventional chucking in an interactions approach here is what we’re saying.

[0:05:32]

 Also, you get many, many terms relating to small pockets of data and so we’re at risk of multiple testing issue, where we find false positives, we find an apparent pattern going on in our sample data and we conclude, right, that must be going on in the world at large, when in reality it’s just some kind of peculiarity in our small data, okay, and we’re doing many, many parameter tests here, many, many coefficients, of course, certain things pop up and so we ideally would be a bit stricter, so that concerns me.

 Okay, so what happens is that researchers, they see this and they baulk, like ugh, let’s pull back, let’s not get so complex. So, one way to pull back the researchers will typically do is that they’ll just kind of focus on one, often two-way interaction of interest, maybe gender by ethnicity, but ignore any potential interactions with age and any other social identities in the model. And this might be theoretically informed, and so contrasts the saturated approach up top. So, that helps make things simpler, it helps reduce the risk of the multiple testing issues because there’s less parameters going on, you’re cutting out the data, that’s fine as well since it’s less noisy, so there are positives in this. But the problem is what if you’re looking for the wrong interaction? Maybe the true interaction is one between ethnicity and age and you’re just not even looking for it now. So, we kind of want to combine the features of both these approaches, in a way. We want to explore everything, which is what the top approach does, but we want to have some kind of simpler way of doing this and some kind of control over multiple testing, which is what the bottom equation does.

 So, that brings us neatly on to MAIDHA. Okay. So, MAIDHA essentially consists of two separate two-level linear regressions, so it’s a direct extension of conventional linear regression to two levels, this is what multilevel modelling is, but it’s getting unusual because whereas in multilevel modelling you’d often think of, hey, I’ve got students nested within schools, hey, I’ve got individuals nested within areas, hey, I’ve got patients nested within hospitals, here the individuals in question are not nested within an institution or a geographic area, they’re nested within their intersections. Okay? And those intersections are defined by combinations of social identities. So, the first intersection might be white male, middle-aged, middle income, and then they cycle through all the permutations and later on, you might have a white female of high education but low income, and later on you get different ethnicities and so on. Each of those is treated as a cluster, as a level 2 unit or grouping.

 Okay, so the two models that you fit, and let’s cycle over all the exciting colours in the palette her, let’s go for a nice pink, so we’re modelling once again the individual health outcome, perhaps BMI, of individual I, in interaction J, okay, and everyone in that intersection is the same as you, they have the same gender, the same ethnicity, they have the same education, the same income, whatever it is. And then we’re regressing individual health outcomes on beta zero, and overall intercept, which will estimate the overall average health across all individuals and all intersections. And then we’re going to have the random part of the model. We’ve got UJ, which is a random effect. Okay? We’re going to capture how the mean outcome for each and every intersection, J, deviates from the overall average beta zero. So, some intersections, just in general who have better health than the overall average, other intersections defined by a different combination of social identities, individuals just have lower health than average. And we’re interested in those differences in average health across the intersections. Those are the inequalities of interest. Now, of course, within a specific intersection, you’ve got lots and lots of people. Yes, they’ve got the same gender as one another, the same education, income and so on, ethnicity, but they’re different people and their individual health will vary around their intersection average, and that’s captured by EIJ, which is the level 1 residual. Okay?

 So, yeah, essentially that’s the first of the two MAIDHA models. So, what we do is we define the intersections. That generates a new variable in our spreadsheet of data. We treat that as the level 2 units or clusters. We fit a two-level regression of health on only an intercept, and what we get back is a breakdown of individual health outcomes into the overall average, an intersection effect and a residual, and crucially those intersectional effects we estimate this new parameter, sigma U square, which is the variant in the mean outcomes across intersections. That parameters quantifies how big are intersectional inequalities. And we put that into perspective by comparing that parameter to the variants of the residuals to the extent to which individuals vary around their averages. So, we’re interested in the relative magnitude of sigma U squared, the variants in the intersection means. And that’s actually going to give us a statistic of the variants partition coefficient, which we’ll mention on the next slide as well, which is like an effect size, how big are intersectional inequalities. It’s a statistic which remains between 0 and 1 and the bigger the proportion then the higher the proportion of variability and individual outcome, health, which lies across intersection means versus within intersections. And so it’s a nice statistic to compare across studies, across countries, or across outcomes and so on, it’s a look(?) outside(?) metric essentially.

[0:11:49]

 Okay, so that’s the first kind of step of the analysis, which is just to quantify the importance of intersectional inequalities by fitting this model and calculating that BPC. Once we’ve done that and established, hey, there are these big intersectional inequalities, we then move on to the more nuanced question of, well, to what extent can those be characterised as simply due to the additive risk factors of the social identities, that in general females have higher health, in general people of lower education have lower health, these additive main effects. Okay? Versus to what extent are there interactions between social identities? So, what’s the relative importance of additive risk factors versus non-additive interaction effects in explaining inequalities. So, let’s do that. I’ll bin my drawings and I’ll jump onto the next slide and let’s have a look.

 Let’s get the pens out, change the colour to what haven’t we had yet, I don’t think we’ve had turquoise, or ice blue as Zoom calls it. Okay, so we have two models. Model once we’ve already fitted, that was the two-level regression of health on only an intercept, where we got out the intersection rounding effect and the residual and from that got the proportional action happening between intersections. Model two is the new one, where we have the same outcome, individual health, but now we regress it on the additive main effects, and in this model we have a dummy variable for sex, so maybe this is the dummy variable for being female versus male. We’ve got three ethnic groups, for example, so we put in two dummies for those to contrast, let’s say, black individuals to the refence category white, perhaps Asian individuals contrasted with the reference category white, and then we have education levels, here you can see mid-ed and high-ed have gone in, so obviously low-ed has been left out, so we’ve got three education levels, and there will be an education gradient in health which will be captured by those betas just like you get in linear regression, and captured the age gradient as well, and so we’ve got low age, which is being the omitted variable, mid age and high age, and we’ll get the age gradient.

 So, from that you can get the kind of additive main effects, the additive risk factors on health of these sociodemographic characteristics. Okay? So, that part of the model highlighted in blue is capturing to what extent I inequality is due to additive effects. The UJ is again the intersection random effect. But now this random effect is picking up conceptually all possible two-way, three-way and four-way interaction effects. Okay? We’ve got four social identities in the model, sex, ethnicity, education and age, so if we were to put in all the interactions in the conventional way, we’d get up to a four-way interaction, but that one term, UJ, is soaking up all of that. Now, every intersection has its own value and we can predict these processes(?), measure and study them, and so it’s a different setup to conventional regression where you have lots more terms in the model and lots more betas.

 So, the UJ now is capturing those interactions, and so the sigma U squared, that parameter is reduced from model one because we’ve stripped out the additive main effects that’s come out of the model, and what’s left in a sigma U squared is the variants of the interactions. So, we can use this magnitude to look at the relative importance of interactions versus main effects. Okay? But we can essentially use it to say what proportion of the initial variability we saw in model one is attributable to interaction effects sigma U squared versus… in model 2, that is, versus the main effects. Okay? So, you can look at the proportion, you can look at the proportion reduction in that parameter across the two models, and that’s the second key statistic which can be compared across studies, if the same study is done in different countries, for example, and then you can compare how that magnitude differs. Maybe interactions are more important in some countries, less in others.

[0:16:21]

 Okay, let’s shift on to the next slide. So, those are the two base models of MAIDHA analysis. Of course, you can do much more, but that is the kind of minimum analysis that people do. And so just to kind of sum that up again, we’ll get out… let’s have a yellow. The first thing is you fit model one, okay, and that allows you to answer how much the intersectional strata matter or how big are inequalities, to what extent is the variation in individual health captured by mean differences across intersections. Okay, that is the VPC statistic, proportion of variants level 2, okay, standard statistical model and modelling, nice effect size. Okay? Then, model 2, we add in the main effects, the additive risk factors, we explain away the inequalities between intersections because we say, well, they’re attributable to these risk factors, but a sum is left over. And the sum which is left over is the part which is attributable to two-way, three-way, four-way interactions. And so we can calculate the second statistic, the PCV, which is to what extent have we explained away the inequalities, and that’s the additive component, and one minus that will be the proportion attributable to interactions.

 So, we get these two key statistics, but then we can go further. Okay? What we can do after that is we can use model 2 to predict the mean health at each and every intersection, and we get those model predictions with confidence intervals, we can do lots of graphical inspection of those mean differences and we can use that to make effectively graphical league tables of inequalities, and we can do nuanced things to say, well, we can do a league table obviously in terms of mean outcomes, but we can do it also in terms of mean outcomes based only on the additive component, or we can rank intersections by the magnitude of the interaction effect. And interaction effects are capturing peculiar, unexpected behaviour that you might miss from a naïve analysis, and so you can see, well, which particular intersections is something peculiar going on different from an additive main effects world and then you can follow that up with further research, and so it’s a nice way of identifying intersections where unexpected things are happening which you’d otherwise miss.

 So, let’s shift on. So, there’s going to be separate R practical and Stata practical video walkthroughs which we’re going to do later on, and you can pick and choose depending on which package you like. Andy will do R, I will do Stata, but it will be the same practical so you can learn by looking at both as well. But here in one slide are the two models. Okay? So, we’ll go for a light purple this time. So, if you’re an R user, what you’d do is you would use Doug Bates’ lme4 package, you’d use the lmer function within that, which is the standard function for doing multilevel models, linear and mixed effects regression, it’s just a multilevel model, you’d regress the dependent variable, health, and you’d regress it on an intercept, which is what the 1 is there, all being very explicit, plus a random intercept across strata. Strata and intersections are synonyms, you can use one or the other in this literature, and so that allows the mean health to vary across the intersections. Then of course you’ve got to reference which data frame the data is in, and it’s a data frame called DF in this example. That’s model 1, and we assign results to the object model 1.

 For model 2, again, lmer, again the same health outcome, again we have an intercept, but now we add in the additive main effects for the social identities which were used to define variable strata, which are sex, ethnicity, and education. And we’ll put those into the model. It knows that those are categorical or factor variables, so will expand it to put in dummy variables for each, leaving one out as a reference category and they’ll be able to estimate their regression coefficients to get the additive main effects. Yeah, and it will again estimate those variance components which then can be compared between model 2 and model 1.

 So, if you’re a Stata user, like me, this is what you do. You use the mixed command or mixed models, which again is a synonym for multilevel models in Stata. The outcome here, well, it’s written actually HbA1c, so I guess ideally we would have written health there just to make these consistent with one another, but, yeah, that’s just a typo on our part. By default it will put in an intercept, and then to make it multilevel, it will add in the stratum, again, it says stratum here and strata there, really ideally we’d write the same word, so strata, stratum, intersection, whatever you like, whatever the name of the variable is, we put that in there and that will give the random effect to allow the mean health to vary across the intersections. Now, the default estimation in R is reml, the default estimation in Stata is ml. Okay? Reml is restricted maximal likelihood, and it’s kind of referred if you have a low number of clusters. In Stata it’s ml, which is a bit more flexible. Just to make the two consistent with one another, we switched on reml as an option in Stata.

[0:22:16]

 And second model, again, the mixed command, again the same outcome, but now we put in sex, ethnicity and education. The I dot just tells data explicitly that each variable there should be treated as a categorical variable and entered as a series of dummy variables or indicator variables, so I for indicator, so just to make that consistent with what we do. And then again we have the stratum random effects, which will now be picking up the two-way, two-way, four-way interactions and again fitted(?) by reml.

 So, we have whole videos on how to do this properly when we walk through obviously more slowly, but just as a high-level one slide how you do it in R and Stata, it’s just like any other multilevel model, okay? The only difference is the clusters are not schools or neighbourhoods, they’re not a pre-existing variable in your dataset, rather it’s a new variable which you generate by the combinations of the particular social identities of interest, sex, ethnicity, education. That is the only difference mechanically in what we’re doing.

 Last slide from me before I sign off this video. You might ask here why not just use loads of dummy variables, the fixed effects approach? You could do a linear regression of health on a separate dummy variable for each and every intersection. Okay? So, if you had 100 intersections, you’d have 100 dummy variables. Now, it turns out that is equivalent to option 2 above, maybe cycle back a few slides, what you’ll see option 2 was when you put a linear regression in between the variables of interest, I think they were sex, ethnicity, education and age, and the two-way interactions and the three-way interactions and the four-way interactions, if you did that, both models fit equally because they’re both saturated, they’re just different parameterisations. Okay? So, this parameterisation described here as separate dummy variable for every intersection is more similar to the parameterisation or the conceptualisation of the MAIDHA model, where we’re directly estimating the mean at every intersection, which is what dummy variables would do as well. But the key difference is, as well as not getting the VPs, the variance partition coefficients and the PCV, those two crucial effect sizes are characterising what’s going on and the magnitude of inequalities in the makeup, we’re also not getting something else, which is that in the multilevel approach, the MAIDHA approach, the predicted intersection means are so-called shrinkage estimates, which pulls the very unreliable sample means for intersections, particularly rare intersections, towards the predicted means from the model based on the additive main effects. And this has preferred statistical properties. You can think of it as a smoothing approach to kind of try and deal with the noise. It also has Bayesian elements, we’re borrowing strength. Okay?

 So, our prediction for intersections 17 is not based just on the data for intersection 17, but it’s dragging in information from across the distribution because these intersections are all related, it’s all part of a system. These aren’t distinct islands and you’re a white female, high income, high education, that that tells you nothing about another intersection, which might be white female low education, low income, because there are things in common here. And that commonality is those additive main effects and so we want to kind of utilise that. Now, chucking in a dummy variable for each and every intersection doesn’t do that, it just treats each tiny pocket of data in isolation and calculates the mean. Okay? It’s learning nothing from the system at large, whereas the MAIDHA approach does.

 So, it sounds quite technical but it’s a really important point. Andy Bell’s got a nice paper in 2019 where he does some simulation studies, which looks at shrinkage also from the perspective of dealing with multiple testing. Clearly, if you have 100 dummy variables and nothing’s going on, 5% of those will pop up as significant on average by chance alone, because we’re testing at 5% level. But in a MAIDHA approach, shrinkage will deal with that, so you then need to kind of turn to Bonferroni corrections, which you might do in a conventional regression. The shrinkage gives an inbuilt conservatism by default, if you like, rather than having to worry about what kind of corrections we need to do to the values. Okay, so lots of nice statistical advantages to the MAIDHA approach, as well as conceptual advantages as well.

 And that’s it from me for this particular video. It’ll be back to Andy next and then you’ll see me later on in this series of videos, so enjoy and good luck with your MAIDHA analysis.

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