# Multilevel models to study intersectionality

## Transcript MAIHDA - Stata example (video 6)

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

Welcome. You chose Stata. Welcome to the practical for the MAIHDA analysis.

So, what we're going to do here is we're going to walk you through a Stata .do file and Stata output for doing MAIHDA. We're going to do this in the context of analysis of diabetes data, in particular a biomarker called HbA1c. Now, this example is actually taken from, I’ll zoom in a bit, a journal article that Andy and myself are some of the co-authors together with Clare Evans, Subramanian and Juan Merlo. And this has appeared recently in, where has it appeared, Social science and Medicine, I think, Population Health, and it's a tutorial paper which talks through MAIHDA in general, but accompanying it are scripts. And what we're doing here is a kind of a bit of a cut down version of those scripts just to kind of keep it simple really for this video.

Okay, so we've provided you with this kind of PDF of all the instructions. You can see the syntax, different Stata commands in there, and there's lots of interpretation and so on and explanation, Six pages, so not so long. But what I'm going to do is I'm going to go through command by command and yeah, speak to this essentially. And so, you can either watch the video and follow along at your own pace, or you know if you don't watch the video, it's fine. Can you just follow the practical independently as well.

So, let’s just send that practical because what we also have is the actual Stata script or do file which we're going to follow, which has all the same commands as the PDF.

So, the first thing you need to do is obviously open up the D file, which we've provided you with. You need to, if you just double click on it then it will set the working directory to that location which is where you’ll also have the data set as well. Prudent to always set the version number to whatever version you wrote the script in, 18.5 in my case. Yeah, if you’ve got an older version of Stata you might need to lower that number or just delete that line, that's worth saying. But more generally, when you're writing Stata scripts, some things can change between versions. For example, the random number generator at some point did, and so you want your work to be reproducible, so you should always really set a version number on any script you're doing. That’s a bit geeky, but yeah.

Okay, so the first thing we're going to do is load the data, and I did that with the use command and the name of the data set which we provided you with, and you can see top right hand corner in the variables window, we've got these variables. HbA1c, a biomarker for diabetes, we have a variable called diabetic which is, we'll see shortly, a binary flag for whether you are diabetic or not. I guess whether you have HbA1c higher than seven threshold potentially or maybe it is actually whether you are being diagnosed as diabetic. Well, you'd have to go to the tutorial article to find the details in full.

But we have sex, race, education, income and age. So, the first thing you should do is always browse the data before you do anything else, you know, just sense check it. Also, you've got 33,000 observations, so it’s a big old data set. So, I'm going to click the browse button, I'll bring it into view. Here we go. And so, here you can see I've got HbA1c, so I’m doing a continuous variable, diabetic was the binary flag. Some people are diabetic and they have higher HbA1c scores than those who aren't, it makes sense. We have sex, which is at least two categories I can see already, male and female. Ethnicity. I can see white, Hispanic, black, so maybe it's three. We'll check this shortly. Education, less than high school, high school, some college, college plus, looks like four categories. It's actually mimicking the original 2018 paper which was a BMI outcome. This is simulated data, but with the same social identities. We’ve got income bands, low, low/mid, mid/high and high. Age bands, I can see 18 to 29, 30 to 44, 45 to 59, I've got 60 plus. There looks like four of those.

Okay, so, these are all numeric values with value labels sitting on top. It's all nice and tidy. You can scroll up and down. It's all well behaved, no missing data and you can scroll down to 33,000, the last respondent in our fictitious data.

Okay. So, that's let's get rid of the data browser for now.

So, the first thing which we're going to do is we're going to generate a new variable, which is this, which is the intersection ID or the stratum, we’re doing to use the language of stratum, these are different strata. So, you can kind of see what's going on actually. If I kind of highlight the five variables which you want to do social identities, this this is always dangerous when George goes off the piece, so let's see if I can break things. Hang on, I want to sort the data. And I'm going to sort it by these five variables. And you'll see straightaway the intersections. Variables must be filled in, so I’ve got to type where I want to sort what’s here, I’ll do it like that. Click, so click submit. Yeah, there we go.

Okay, so I've sorted the data by the social identities. You can see, we've got a block of people who are identical sets, identical ethnicity, identical education, identical income, identical age. They have all of this in common because they are a common group and so we call them maybe Intersection 1. Again, if you cycle through, at some point here we cycle through and age now changes. So, now gender, ethnicity, education and income, the same as before, but age is now, we’ve shifted from young to middle age, 30-44. So, that's good. I'm just about hanging on into that age category, so that's pleasing. Yeah, and so on. So, you can see the intersections already. So, we need to generate a new variable which takes a constant value for all individuals in that first group and then increments from one to two the next group and so on.

So, let's see different ways of doing that. Let's see how we're going to do it. This is quite fun. So, we're going to make a new variable called Stratum, which is equal to clearly a function of sex, race, education, income. Triple line just means we're spilling over to the second line because I can't fit it all in one. You know, if you don't like that, you could do this. You know, that's fine as well. But then I go over the kind of the standard 80 character width they use in scripts. So, people would often do it like that, the line continuation symbol maybe at the end of that.

Okay, so what we do is because sex is coded, I don't know whether I can, this is something I always fail to find. Where’s this? I want to kind of switch off these value labels. How do I do this?

I'd quite like to not look at the value labels temporarily and, yeah, data utilities, label utility. No, don't know how to do that. But you can switch off, you can toggle between seeing these words and the underlying numbers. If I just click on here, click on no, this is at number one, and of course sorted the data rather unhelpfully. Let's just sort it fairly randomly by the outcome. And only partially variability because of the systematic structure in that problem. If I click on a females code 2, yeah, white’s coded 1, Hispanic, black with 2, Hispanic 3, income’s going to be 1, 2, 3, 4 for income bands. Age bands are going to be 4 for 60 plus, the oldest group, 1 for 18 to 29 and so on.

So, they're each kind of coded by an integer which just goes 1, 12, 123, 1234, depending on how many groups there are. So, then we can do magic maths, yeah? If we go 10,000 times sex, okay, you're either going to be 10,001 or 10,002, if you then add on 1,000, so 10,000 x 6 is either going to be 10,000 if your sex is 1 or it's going to be 20,000 if your sex is 2. So, we now add on 1,000 times race, we're going to increment the second digit of this to be 1, 2 or 3, okay? And then if we add on 100 times education, we're going to increment the third digit of this to be education levels 1,2 , 3, 4 or however many there are, and then 10 times income is going to increment to last and then we finally add on one times which age band you're in, 1, 2 or 3 or 4, and that's going to make the last digit 1, 2, 3 or 4. Which is kind of neat. So, let’s do that.

And then here straightaway, not only could I generate the variable, which you can see in the data browser, we have also, I've also tabulated this, got a lot of these, right? I've tabulated these as well and you can see the first group is 1, 1, 1, 1, 1. That's in the first category of each of the five social identities, and it's 41 people in there and so on. So, that tabulation is quite useful because it tells us the percentage of people at each intersection, and because I've got many intersections, okay, I’ve got lots of these, maybe I can even do something like code stratum. If I do that, it tells me I’ve got 384 unique intersections, okay. Unique values of stratum.

I might want to make a variable in my data set which holds the number of people in each stratum. So, if I sort by stratum, you'll see we're back to the male whites, low education, low income, young. They are coded by five 1s. So, how many people are in this group? I can see 41 are. So, let's generate a new variable which will store the numeric value 41 participants in that intersection here and then however many for the next one. So, we do that by saying by sort stratum to sort the data, which I've just actually done, okay. And then within each of those generate a new variable called strataN which is equal to \_N, which means the number of observations in each strata.

So, if I do that, let's get rid of this sort window, and if I, and you'll see it pop up here, okay. So, I'm going to go like this, I'm going to run it and you'll see the variable pop, which is satisfying. Pop. There you go. That's the 41, okay? And the next one, which was exactly the same but now people are aged 30 to 44, that's got 55 individuals. Nice.

So, let's do some descriptive statistics. Probably would have been good to do the descriptors before generating the stratum really should be the first thing you do. So, let’s tabulate by sex and we can say about, pushing on 60% are female. By ethnicity, white black, Hispanic, about 6% white, 20% black, 20% Hispanic. It's almost like we've simulated the data, which we have. Tabulate education. We've got 15%, pushing on 30%, 20%, 35% for those four education bands. Income. They look, yeah, 20%, maybe around 25%, pushing on 30%, 30%, 20% for these four income bands. Age. 15% are the young, 30% young to middle-aged, middle-aged to a little bit more, a little bit older, 30%, and then 60 plus is that 30%.

Yeah, so it’s all kind of pretty nothing too extreme there actually.

Here's the outcome. HbA1c. Now that can range in this data from a low of 10.3 to 101.17 with a mean of 40, a standard deviation of about 10. I'm not very knowledgeable on HBA1C, it's not telling me anything substantively interesting about that, but I can show you a histogram. Always important to do histograms in your continuous variables. Where is it? It's going to pop up at some very slowly. There we go. Pop. First graph, you try to attempt to plot an early stage session it always takes longer because it thinks about graphs, subsequent ones are quicker. This is a, what odd shape distribution, right? You know, kind of goes up, looks like a normal distribution until you get to about the mean, but then it kind of comes down kind of like excessively sharply. Then it has this long tail, okay. So, we are working with big samples here. You know, we've got many clusters, many individuals per cluster. But even so, this non-normality would transformed through to the random effect residuals, for which we make normal assumptions. So, those might be things you might check out later. They'll always be less problematic in the preferred Model 2 when we start putting in additive main effect, which is part of the generation of the nature of this distribution. It's not so bad in the big scheme of things.

Okay, let's move the graph out the way for now. I'll put it back later. Because we're going to fit our first model, okay? Now, this is the MT2 level model of the continuous outcome HbA1c. So, the command is mixed, the outcome variables is HbA1c. No explanatory variables. And then we have these two vertical pipes at the start of the random part of the model specification. And include a level 1 residual, an individual residual by default to make it a model. But what is optional is whether you include a random text to turn this from a single level to a two level model. We want that. And we define the clusters by strata which again is that kind of new variable we've generated, which is our cluster idea level 2. We fed the model by rem, or restricted maximum likelihood to override the default of maximum likelihood which, yeah, it's a little bit better, I think, for propagating uncertainty into the predictions, you know, it makes less and less of a difference the more clusters you have. We've actually got a lot of clusters, so not so important. But at least it matches the default in R which is what we're trying to replicate, the R script was written first here. We're trying to replicate that in the Stata to some extent. So, to be consistent between the packages will hit reml on. So, there we go. Here it's estimating, there's maximum likelihood, so it's rates to convergence and we see some output.

So, what we've got here was confirmed with fitter in a mixed effects model by reml estimation, it's confirmed that the clusters are the stratums that I don't identify, the strata. 384 strata. There are 33,000 individuals in total. We've got at least one group or intersection which has just one person, okay, on average we've got 86 individuals per intersection. And the biggest, you can imagine that being a big group like white middle-aged, middle income, middle education majority, majority, majority type group, that has 683 people.

To fit the model by restricted maximum likelihood. So, the log restricted likelihood is printed out at the end of convergence. Then we get the parameter estimates. Here, the first table here is the regression coefficient for betas, the second one of the variance components. In terms of the betas, we've only got an interception, so this is estimating the average HbA1c across all intersection individuals, about 41. And that's not particularly interesting in its own right. What is interesting, however, is the breakdown of variance here, okay? We've got, what's that, about 9 plus 90, that's basically 100. That's sums to 100. So, the add up the variance components to get the variance of the outcome, which is 100 and of that variance ten, or 10%, is between the intersections and 90 is within. So, you can even do a kind of is that ICC and that will give you that proportion of 0.10 or for precision 0.094, 9.4% of the variability in the HbA1c is across intersections. So, that's the magnitude of the intersection inequalities with respect to, what was it, sex, race, education, income and age. So, that's kind fairly in line with the literature.

So, next up we're going to store the model results because we might want to pull them back later. We're going to predict a new variable and we like to see these things in action. So, if I bring back the old rows and make it smaller, scroll to the right, get an empty, we're going to populate this column here, okay, in our spreadsheet and we're going to populate it with predicted HbA1c values, the mean HbA1c predicted by the model in each and every intersection and these are shrunken. Pop. Again, then that first cluster, they're predicted 37.4, which is lower than the overall average.

Okay, then we're going to fit the second model. Now, the second model, the only difference is the bit I'm highlighting, which is I now put in the active main effects of sex, race, education, income and age, five categorical variables, tell Stata that they're categorical, so don't treat them by default as continuous by doing i.. Now, i. will add in indicator variables which are dummy variables and then leave one out for each variable. So, I can run that, again, maximum likelihood or restricted iteration, iterates the convergence.

So, I've got some results here, so I might switch on some options. If I type mixed again, I can print the results again, but I want to print them different. I'm going to write base category. There, George, this is where it’s difficult, so not difficult, I just can't handle the option. So, I go help mix and get the options. I want to see what the reference categories are. I want to make that explicit. So, there's some display options somewhere. Display options. Display options, base levels. There we go. Get out of there. So, it wasn't base category, it's base levels.

Okay, so now it's said what the base is. The male was the omitted dummy, the female dummy was put in, females relative to the males score lower on HbA1c, they have lower biomarker for diabetes is less, less diabetic if you like. But black and Hispanics versus the base of white are both higher and significantly so. It's got the education gradient you're getting. You're getting basically more negatives. So, the more educated you are, the lower you score and the better health you have. The higher income, no significant differences. So, low to mid and mid to high income are not significantly different in their mean HbA1c levels to the low income. But the high have significantly lower HbA1c.

Now age gradient is a positive and significant, and intercept obviously relates to someone who's base male, white, low education, low income and youngest age. That was intersection 1, 1, 1, 1, 1. Okay? But crucially, that is a prediction based on out of the main effects for intersection. Even that intersection might deviate from that and this variance here is the variance in the intersection interactions, which is how every intersection, 384 of them, deviate from their predicted mean based only on active main effects and every intersection may be above or below. Half will be above, half will be below the sense of these interactions are zero. But even with those interactions included, we've still got this kind of variability which we haven't explained at all because this is within intersection variability and these are all level 2 variables, so they can't explain the level 1 variance there.

So, let's store those results. Let's shut the predictions as another variable. Now, crucially, the predictions from Model 2 differ slightly from the predictions from Model 1, okay. Why is this? Well, the prediction for Model 1 has shrunk towards the national average, and we're observing those shrunken predictions, whereas in Model 2 the same observed data, the same observed mean in every intersection, has now shrunk not towards the global average, but is shrunk towards the prediction for that intersection based on the additive risk factors which we've added into Model 2. And these predictions would typically be better, more accurate if you like, than those, more trustworthy. Okay.

So, let's kind of also predict the interactions with Model 2. I didn't show you them pop up that time, but then another column in here, make that a bit wider. Okay. So, here are the interactions with our standard errors. Now these interactions are probably better termed as deviations from expectations. But this is how the predicted mean, which we’ve got there, deviates from what we expected, which could be another column again based on the main effect. So, that first intersection, we're predicting them 37.65 and that is 0.377 of a unit less than what we expected them to show given on the global patterns. Okay? So, but the standard error is big on that and actually bigger than the deviation. So, that's not significantly different. Okay? So, any deviation from expectations for that first group are not to be over interpreted at all because it's small and not significant.

Okay, so that's kind of, let's collapse the data down. We can do this quite fun as well, to one record stratum. So, the new data set will have 384 rows. Whoosh. Okay. And we've done that here. And you can see, I might even sort, data sort, sort data on that. Okay, hang on. Oh, yeah, there we go, yeah. Maybe it was sorted anyway, I just wasn't at the top. Okay. So, that's the intersection. So, you can see now I've got one red intersection, we've got 1, 1, 1, 1, 1 and then increments so, 11, 112, 11, 113 and so on and we can just cycle through them. And of course, what we're doing is we're cycling over the combinations of these five social identities. And we've got the number of individuals per intersection. We've got the predicted mean from Model 1, the improved predicted mean from Model 2. And in terms of Model 2, we can see, well, to what extent was that made up of an interaction, a deviation from expectations, and then was that significant judged by its precision? And then we've even got here, this one here is the observed mean. So, we've got the observed mean, the predicted mean from Model 1 and the predicted mean from Model 2. The observed mean is the worst because we've only got two people here, okay, so it's all over the place. The prediction of Model 1 will be better than that, but the prediction from Model 2 would be the most trustworthy because we’re borrowing strength beyond the two people across the data.

So, yeah, let's have a look at some of the commands here. So, yeah, so when we wanted the predicted means, we just used the predict command, the name of the new variable and the options fitted, who are fitting the data. If I wanted the predicted random effects for Model 2, we just did it in Model 2, so we predict new variable name, we are RE, R effects, RE, random effects, you know, random effects and the ram effects, SES standard errors are stored in another variable name. So, that's how we got the standard errors which we'll need. And then we collapse the data down, members of one record for intersection using the collapse command. We collapse by what, by intersection, and we list any intersection level variables here, which are basically all of them, so we can retain them in the new collapse data set. And so, really all we're doing is collapsing or averaging mean of the individual HbA1c values and so now when we have this transformed data set, that HbA1c is a mean balance.

Okay, so let’s compare Models 1 and 2 in a table. Here we go. And so, what we see here is this is Model 1, this is Model 2. Model 1's only got an intercept in. We've got the variance. And this was 10 versus 90, so 10% of the variance was between the intersection means, you can improve the labelling of this stuff. Once we've chucked in the added domain effects this drop from basically ten to one, so, I’ve explain that basically 90% of the inequalities has been explained away by the addition of additive main effects. We went through their coefficients already. But we can't explain within an intersection why some people have higher scores than others by the variables we could put in, so that's unchanged.

So, we've done Models 1 and 2, we’ve talked about the P, the VPC, which is basically 10% of the variability was due to intersections and of that 10% we can explain 90% approximately by additive main effects, 10% is unexplained interaction action.

So, let's try and identify which particular intersections and this is the last section of the script, are doing the work. So, what we want to do is based is focusing on Model 2 now which we're calling, we called it a, what did we call it, A and B, M1A and M1B. So, A and B, Models A and B, the empty models. So, in terms of Model B, we can rank them, okay, and let’s sort the data as well by that. And you can see extender generate, new variable name, function rank, the old variable and so you can see, this just has these in rank order, okay. And then I can plot a caterpillar plot, so a standard error bar plot, plotting the means and the standard errors, and they're going to turn standard errors into confidence intervals by scaling them up by 1.96. On the X axis we're going to rank them from using this variable from 1 to 384. We're going to draw a horizontal line at zero. We're interested in how these, well, that's not so useful actually. Let’s have a look.

Yeah, that Y line at zero is not too helpful. It'll be better if I kind of stuck in like a line and at 41 you type in the exact mean estimated by the model, and then you can see to what extent our intersections have, that's the horizontal black line, which has gone in. So, that's an improvement, changing that from zero to 41. You can see how the predictions for any given intersection deviates and whether significantly so from the overall average. So, Model 2 predictions, the most trustworthy of the three sets, they're kind of distributed. We've got a bit of a kind of, this is kind of reflecting the histogram earlier which had that kind of positive skew. That's this kind of, these intersections here really. But a lot of statistical separation. And these are the means. They account for about 10% variability.

So, let's sort the data. I think it was already sorted, and the reason why we're sorting it is because we’re going to list in the window the first six observations. I think I'm going to have to make the window bit wider at this point. So, let’s anticipate that and go whoosh, that’s quite fun. Yeah. So, we'll do, we'll list them now. Still too wide. Okay, so let's get rid of all this stuff here. Try again. Yeah, there we go.

So, this is the league table. So, because I sorted the data form, well, it's actually sorted from lowest to highest. We're listing the sixth strata with the lowest means. Okay, these are the lowest means. Okay, so who's at the bottom of the league table? It's dominated by females, so females have lower HbA1c. They’ve got better health with respect to this biomarker of diabetes than men. White is dominant there as well. Education is not, doesn't matter so much, but higher education looks desirable and higher income tend to have lower HbA1c as well, and young. So, the young females who are white, who are more educated and higher income, they in general have the best health in respect to this. You can guess what the opposite will be. Males. And now it's black, okay, so whites were at the one end of the distribution, blacks were predominately the other or more often than not, at least, certainly in the top six. Education levels are now lower because we've got these higher HbA1c. Income levels are lower and it's in particular it's the 60 plus. Okay, so that's some league table work. We might then rank the interactions which are stored in, you remember this Model B, the random effects, the second model, the random effects are predicted interactions of deviations of the means from expectations, you're going to rank those. And we're going to do now a caterpillar plot of, let's bring it into focus, of how intersections deviate from what we expect.

So, here this advice, it goes to the scale, you know, again from about -2 to +2, we've got a distribution, there's a few intersections at the right top of the distribution whose predicted mean, and these are shrunken and conservative means, for the mean HbA1c for those intersections. What we're saying here is that the means that we're getting are higher than what we expect by main effects because we've got this positive deviation, whereas the means down here are lower than what we expect given the additive effects. Okay, so it's quite interesting to follow and there are a few significant ones. It would be quite interesting to see who they are and to try and think, well, why are certain combinations of characteristics leading to better outcomes than what we might otherwise expect? Or worse outcomes than what we might otherwise expect by the naïve main effects analysis.

So, I'm going to generate a new variable. Let’s bring them back to the middle. And I’m going to generate a new variable. Let's bring this one back as well. Whoosh. Which is to generate a new variable called SIG and these brackets, like it's going to evaluate this to true or false and if true assign a one, if false assign a zero. So, SIG will be a dummy variable and when will it be evaluated as true? We're got an all(?) operator in there. This is kind of Boolean algebra. So, either the clause before the operator or the clause afterwards is true. Then this is a single split back of one. But if they're both false, it will float back to zero.

So, basically what we're trying to do is we were trying to work out which of these are significant, okay. So, if, starting at the left, if these guys here, if the top limits are less than zero, okay, which is what we've done there, then it's significant. Okay? So, what we've done is we've taken the dot, which is the predicted interaction effect, plus 1.96 times its standard error to get its upper limit of its confidence interval and if that value there is less than zero, as it is what I'm highlighting, that will evaluate is true and then we're satisfied and SIG will return to one. Okay? But if it's over, if it's above, then that won't be true, and then we need to look at the second half of this, but it might be the case, like for example, the other end now is saying it's the lower limit is the dot -1.96 times the standard error. It's a lower limit above zero. Okay, and so this is kind of capturing whether you are either the upper limit of below zero because one of these, or the lower limit is above zero if you're one of those, and focus isn’t significant.

So, whoosh, that was quite tiring. So, we're going to kind of generate that variable. Here's a binary flag, and here's a significant, this is significant, that’s significant, there are not many of them. Let's just keep the ones which are significant. Okay. And there's only 12 of them. And let's rank these and let's rank them by, so make a new variable and rank them by this guy here, okay. So, there we go, we’ve ranked them. We might sort the data just so we can visualise it, not visualise it, so we can see it in the browser.

Okay, so we've ranked them from, and this is a subset of interactive vectors which are significant, half of them will be significant negative, actually more than half, eight. So, the eight at the bottom of the caterpillar plot and then four at the top. Again, they're the most extreme, well, typically the most extreme. They've got to kind of fairly sizeable as they are, they've got to have enough data to support a tight enough standard error to have a competence interval, which excludes zero and so all of them do. So, we don't have any super rare categories. And we did a caterpillar plot. And so, you can see now I've got the subset of what I had before. I've got the eight which are intersections which have unexpectedly low HbA1c and they're healthier than what we'd otherwise expect. And then we've got four which are less healthy than we’d expect. And then we're going to kind of, we're going to pimp the plot by adding some options here. So, that’s the intersection IDs. So, let's just look at the kind of, this is a bit new. What I did was I got exactly what we had before, including the first instance when we did this, but I've got the line continuation and then I've added another plot on top. I said add a plot which is a scatter plot which can give me the green dots. Again, it's got the same, the predicted interaction effects on the Y axis. Again, it's got the rank on the X axis and for that scatter plot I've got some options like continuation which are the end label, the marker label. So, these are called markers in Stata parlance. We're going to label them with stratum ID, which is the stratum variable here. So, these numbers or IDs will appear on plot if they do appear. And we can control where they appear in the plot by saying mlab, marker label position, 12.00, top of the clock. Okay, so that means immediately above the marker, but then we don't want them just hanging around here. We want to push them up the page so it looks a bit neater and that's the marker label gap and you just play around with the number here, right? If I change that to two, they're going to drop down the page, which is a bit rubbish. Yeah, they're all overlapping now. If I change it to 40, they'll be too high up the page. See just give it a try. Do high up the page. I said 20. Okay. And then because I've got two plots on there, an error bar plot and a scatter plot, it wanted to kind of, well, it would add a legend by default, take the legend off, option off by default. That sticks in this. And that's all a bit of a mess. So, I'm going to stay legend off. And there we go. Nice and neat. We're done. Super-duper, everyone. Well done.

So, what we did there was we visited the tutorial data set. Here we go. You get all the pointy click instructions and lots of explanation here. With the tutorial data set accompanying the tutorial paper we wrote recently in 2024 with Clare Evans and co-authors in Social Science and Medicine, we took that data set which has even more advanced scripts on the journal Web page to look at, and we did a simplified version of it here with a talk over. We basically fitted MAIHDA models to the continuous outcome twice, the empty model, model with additive main effects. We did that to establish how big are the intersection inequalities in HbA1c? 10%. To what extent is that explained by additive main effects? 90%. To what extent is it therefore attributable to interactions of two-way, three-way, four-way, five-way conceptually? The remaining 10%. Can we identify any specific interactions precisely enough to make a statement about specific intersections, which is a much bolder thing to do? Yes, we can. How many? 12 of them, eight intersection were identified as being unusually low HbA1c or unusually high, unusual with respect to what's going on. And of course, what you can do is you can go and look at who are the eight, okay, with these, where are they? With the negative, significant negative interactions. And we can see what are they, is there some commonality, what's going on are there particular, and commonality here would be in terms of pairings. Pairings. So, we're seeing kind of, we're seeing maybe female and white, okay, one, two, three, four. Female and white are showing unexpectedly low. HbA1c is what that's saying there, whereas we've got a bit of all incomes in there, a bit of all ages. So, that looks like the kind of ethnicity by gender, looks like a kind of a hidden two-way interaction of play potentially.

Anyway, I think I've done enough talking away, yakking away. So, well done for following this mini course in general and the Stata practical specifically. And good luck with your own MAIHDA analysis. There's lots of other materials on the web to support you to do MAIHDA, so do have a good look around, other practical video tutorials and so on. More and more stuff coming on, along with loads of applied papers to scintillate your interest in this.

Thank you very much. Bye bye everyone.

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