Cross-Classified Models, Part 4-Practical

Hi good afternoon everybody, my name is Professor Bill Browne and this is the practical using MLwiN that goes with our learning materials on cross classified models so you should have watched three videos 2 by myself and one by Professor George Leckie on cross classified models. And in those videos we covered a particular practical example, which is the Fife data set from Scotland and in this talk I'm just going to go through taking you through the practical that goes with it, so you can see, on the screen here Chapter 15.

This is my MCMC manual for the MLwiN software and this chapter, this goes through cross classified models so there, the first couple of pages, as I scroll through here cover a lot of the material that we've covered in the lectures. They give definitions of things like classifications, they talk us through notation and they even give us a little bit of background on the Fife dataset.

So I'm not going to scroll through this all the time. What I'm gonna do is I'm going to start up the MLwiN software I'm gonna work my way through this practical and you can watch. So if we go on our start menu "Center for Multi level Modelling" there should be MLwiN version 3.05, which is the current version. So I click on that. Any minute now the MLwiN software will appear.

Okay it's slowly starting up here, I think, because I'm recording at the same time, and here, you will see the MLwiN software, hopefully you can see that on the screen. So I'm not going to show you the text, while I do this. I'm just going to follow the instructions So the first thing we've brought up MLwiN and the first thing we need to do is, we need to open up the worksheet.

So in the In the instructions that if you have for that chapter on cross classified models they will, in boxes they will tell you what to do and I'm just following those there so we can open the sample worksheet And the list that appears we choose xc1.ws, which is the Fife dataset in Scotland and when you open it up what it does, is it creates this names window here, which gives you a list from the top to the bottom of the 11 variables in this data set from Scotland.

It talks a little bit about those different variables you will have come across them if you've looked at the lectures so let's go straight on and choose from the data manipulation menu, "View or edit data", and that gives you all of the data and it specifically asks you in the instructions to choose some columns so "attainment" "Primary ID", "secondary ID", "pupil", so in MLwiN it's basically just shortened to four columns.

Here's the four variables, so what we see here I've moved it across to the right hand side of the screen here the attainment scores, so the first individual go an attainment score of 2, and that individual is in the first primary school primary school one and secondary school one and then pupil: they are Pupil 39. Pupil ID is a kind of an arbitrary number in some respects and we come down here, the first 8 are all in primary school one, and then we have some in primary school five and so on.

You'll think well is this data nested or is it cross classified, so if we follow the instructions, if we were to go to line 1355. Here we see some more primary school 1 so there's some individual pupils in primary school one who went to secondary one, but these ones here went to secondary nine.

So this shows you that we have a cross classified structure. So the first thing we do next is we try and set up some equations to fit models, so we go to the Model menu, we choose Equations, and then there is the.

The equations window, you can see there and I follow the instructions, the first thing I do is click on Notation and because we're doing a cross classified model we get rid of what we call the multiple sub scripts. Done, and you can see that we just get now an i in subscripts for our response variable, y_i.

So I'm going to set up my model. What I'm interested in is these attainment scores at the children at age 16 so I choose attain as the y variable. I choose 3 classifications I'm going to choose them in the order of secondary school as classification 3, Primary School as classification 2, and pupil classification 1, so I click Done that's that bit done and now we need to specify the right hand side of the model, so I click on the red X_0 and I choose cons which stands for constant it's just a column of ones and I'm going to say, to start with let's just have a model with secondary school and pupil and I click done. I can click the estimates button, a couple of times here we have a full description of the model.

Okay, and if I follow the instructions, if I click Start this will fit that model in a classical framework. We're using MCMC for much of what we're doing so, if I go to estimation methods and click MCMC this gives me the different things I can change there. I'm going to keep them as default and I am going to click to close that actually I then click start and away will run the MCMC you'll see the numbers are changing. Actually if I get rid of some of these windows it will go quicker.

Okay, and it's finished down here we can see actual updates 5000 so that's run for 5000 iterations. I click plus to get a full model description here so we have a model Okay with 3435 observations 19 secondary schools 303 primary schools. Which you might think well that's a bit odd as the actual data has less than that. And what this has done at the moment we haven't fitted primary school in the model anyway but it's nested primary schools within secondary schools so every time it comes across a unique ID it will call it a new primary school, but ignore that for the minute.

At the moment what we've done is we fitted the model with just secondary school effects here's the model Okay, we can see the average attainment score is 5.6. There is some variability due to secondary schools and some unexplained variability at the bottom, and we can use Model-> MCMC -> DIC diagnostic to give us some measure of how well that fits. so there, we can see how the DIC criterion is 17309.99 and you'll find that number in the chapter as well.

If I was to refit that model without those secondary school effects, I can do that by changing back to the IGLS algorithm the classical way of fitting them and just coming back into cons and removing secondary school for a minute. Click start. The model needs MCMC so I've changed estimation to MCMC and click on Start.

Okay, and that model fitted almost instantaneously so here's a model. That model just has the attainment score and some variability. We've got no levels in the model, I can get the DIC diagnostic for that 17431 so we can see that that has dropped when we fit the model with secondary schools and so that's a better model.

So we find that actually we do need to include secondary schools in the model but that's just the nested model what we're interested in is a cross classified model, so if I go back in again I change back to IGLS so every time when we are fitting MCMC, we have to go back to the classical methods

just simply to change things in the model simply because we need to get new starting values for our MCMC methods, so I click on beta zero okay put secondary back in and following the instructions here I click on PID(2) and now I have this structure. We have, if you look here, the attainment score is some constant, some effect for secondary schools and some effect for primary school and some Level one residual. So there's where we have got to now.

If I was to look at - So if I was to go Model-> hierarchy viewer we can see quite clearly that in this model it believes that there are 19 secondary schools and 303 primary schools, but actually we know in the dataset there is only 148 so every time it comes across a new id for primary school it believes that that's a new primary school so every time, so the pupils who go to secondary school one and primary school one that's for those primary school, one then the pupils who go to secondary school nine and the primary school one that's thought of as a different primary school one so that's why this number is 303. So we have to tell the software - so if I run the model first with IGLS OK, and then, if I go in and go estimation methods MCMC. I can go in and go MCMC classifications, and now I tell the model that actually my data is cross classified. I can come out and I click on done and lots of things disappear.

And now, if I bring up the hierarchy viewer we get a different thing here which talks about the different classifications, and now it's correctly identified that there are 148 primary schools. And if I come down there in the equation window, we also see that there are 148 primary schools. I click Start that will fit that model using MCMC and what we will see now is that we will get different variabilities - so we can see actually primary schools explain much more of the variability then secondary schools about three times as much here 1.2 nearly to 0.4 and is that a better model?

Well let's first off just check out the model. Let's look at the trajectories plot because that's what they asked us to do in the manual. So if I go to trajectories, this is just showing me how well the MCMC is doing. Look at the secondary school there, and so I just click on that trajectory plot Okay, there is a plot okay, so we can see really that there is some variability and that is a reasonably well converging chain, we have got quite a skewed distribution that's because we only have 19 secondary schools.

Okay, so if we were to now just say, well, is this a better model, so we can do that, via the DIC again so if we go MCMC DIC diagnostic now we can see that the DIC has dropped to 17047 compared to 17309 so that's a better model now okay so we've got a better fitting model.

We can look at other things in that model, we cannot get residuals so if we go Model -> residuals. We can choose the Level at which to look at the residuals, so if I look at Level three, for example, click on the Calculate button and if I go to the plots I can look at residuals versus ranks and if I click Apply here is a diagram so you can see here a triangle for each of the 19 secondary schools and they are sorted from the worst performing to the best from left to right and this one, the worst performing, looks a little bit of an outlier. It's got very low residual I click on it, it tells me actually that's school 19.

Okay, so we'll come and look at that a bit more later in this practical. We can also get the primary school effects so if I go back on the residuals to the settings tab I can change from secondary ID to primary ID then Calc and lots and Apply. Here we have the graph of the 148 primary ids, a triangle for each primary school and we don't really have any evidence so much for any outliers, so we can click on, for example, the lowest one here. We can see that that's primary ID 139, for example, so we

can do more things like that. What about having a slightly more exciting model? So if I close up some of these windows, let's close up the residuals for now.

I'll go back to IGLS so I can change the model. I'm going to add a term so I'm going to make that the attainment scores depend on vrq, so vrq is a kind of verbal reasoning test that was done at the.

start of schooling. So, I can start that to give me some starting values and then I can change to MCMC and it remembers that we're doing a cross classified model, so it immediately goes to 148 primary schools. I click on Start and you can see the counter ticking away in the bottom corner. Okay, and that's finished. We're currently we're not getting too involved in how long should we run for? we just going by the defaults. That's not what you should do in practice, but in this case it doesn't matter too much. In practice, you should really just check that your model has converged properly.

So here we can see the model we've got a strong positive effect, 0.160 here for verbal reasoning score so for every one extra verbal reasoning score the attainment goes up by one ,no I mean up by 0.160. sorry. So we can say is that a better model? Well we do the Dic again and they're all coming up here and the DIC has dropped quite dramatically there so the VRQ that's really explained a lot of the variability. OK, and now what we can see is really we have of what's left about 6% is here, this value here, 0.278 or 6% of the remaining variability is due to primaries and only about 0.4 is due to secondaries.

So I could carry on like this, I could try all of the different, different predictors. I'm going to do them in one foul swoop. I'm going to add in several terms. I'm going to add in social class and click Add Term again and I'm going to add in father's education. I'm going to add in mother's education and finally I'm going to add in choice, which stands for school choice as to whether this is the first choice for the child or not. I'll start that using IGLS to get starting values. I'll change to MCMC and I'll Start again and again I've got a slightly more complicated model so it will chug away for a little bit.

Okay, and then it's finished now, so we can look at these predictors and we've seen this in George's lecture verbal reading test is still very high. Very significant effect. Social class, the higher the social class the higher the attainment. The higher the father's education the higher attainment, the higher the mothers, education, the higher attainment and if you don't get your first choice school so school choice is 1 for first, 2 for Second, I think that somebody was very unlucky and got fourth maybe then you get a negative effect so not going to the school you want to has a negative effect okay so that's interesting.

We can look at the DIC for that as well, so if I do MCMC -> DIC diagnostic and we have dropped again we've dropped not quite as dramatic a drop as when we added verbal reasoning test but we've dropped from 14805 to 14728.

Okay. We can look at the residuals again in this model now with these additional terms in it, so if we go to level 3 calculate and if I plot that graph of the residuals - again Apply, we can really see this drop here between this school, which is still school 19 and the rest, so the question really is looking at this, this model here and really is that little bit of variability left in secondary schools really just a difference between school 19 and the rest?

So what can we do about that? well if I go under data manipulation is a command interface window and in here, I can type a command. I'm going to type calculate c12 so it's a new column is equal to school ID in quotes double equals 19 and what that's going to do is generate a column.

That seemed to work okay and I am going to name c12, quotes again, school 19. Okay, so I could look in the data and if I view data I can see that school 19 is there, what we find if I drag this across you can see lots of zeros and at some point we'll get to school 19, should be right at the bottom, we've got 1s so it's a dummy variable for school 19.

Okay, so we're going to use that to compare the model we currently have to a model where we just put in a dummy for school 19 and don't put in secondary school. So let's change the model to look like that, so I go back to IGLS I click on the cons here and I'm going to remove the school IDs. Okay, so they've vanished. Done, and then I Add Term to add in school_19 so rather than having a random effect for each school what I have now is just a fixed effect for school 19 Okay, so I click start, get some starting values, change back to MCMC OK and click the Start button.

What you notice in MLwiN is that windows appear to jump around when you press buttons! So there we go that's finished. So there we can see the output, the windows and what do we see here, well we can see School 19 indeed has a strong negative effect - we were expecting that, we don't see much difference to the other effects here and the variance 0.209 for primary schools is very close to the point 0.206 that we had previously. Okay, in actual fact, that the lower level variance is 4.168 and was 4.171 before so we've not really changed much at all.

If we look at the DIC diagnostic what do we see? Well actually if we compare this model with the last one it's actually moved down ever so slightly so we've got a slightly better model. So really what have we done there? We've said that we started off with a cross classified structure, and it was really important to have both secondary schools and primary schools in it. Then we've added in these various predictor variables and when we've done that that's been enough to explain some of the differences between secondary schools and in actual fact what we were left with is really just a discrepancy between one outlying secondary school, school 19 and the rest, so if we instead just use a dummy fixed effect for that school. We get a better fitting model, and now we no longer need to do a cross classified model, which means actually we wouldn't have to do the MCMC estimation, that we do we could go for the standard IGLS algorithm.

So I think that's all I'd like to say here -that's the end of the chapter. So hopefully that shows you a little bit of how we can do cross classified models in MLwiN, so thank you very much for listening and I hope you can follow that yourself at home with thechapter.

Goodbye.