Cross-Classified Models, Part 3- Extensions and Further Applications

Hello, and welcome back to Part three of this series of lectures on cross classified models, my name is Professor Bill Browne and I co-direct the Multi-level modelling Centre and in this part I'm going to look at extensions to the standard cross classified model which we've seen where we have 2 higher classifications and also extending out into other applications.

So let's have a summary of where we've got to so far, so in in the first of the three lectures I start off and I introduced the idea of cross classified models and kind of motivated that by talking a little bit about multi level models in general, and particularly nested cases, and then showed what the difference was with crossed classified models from some illustrative examples and gave some background into a practical example that we will be using in a practical that you can follow, along with, but also that was covered in the lecture. And then lecture 2 my colleague, George Leckie took over, and he talked about how we fit cross classified models in practice.

In this lecture he also introduced some notation for cross classified models and some useful diagrams called classification diagrams and I will exploit both the notation and diagrams in this lecture that we look at here, so in this lecture really we kind of look at 2 further examples. Okay, and he's going to illustrate extension, so the example we looked at this practical example that the five data set that we've looked at so far. That data Center has to hire classifications has primary schools and secondary schools and it has a response variable that we assume is normally distributed Okay, in reality, it actually takes values, one to 10 and integer values but we're assuming we can fit a normal response model to it, in other words the residuals will look normally distributed and they do, to a certain degree, as well we're gonna we're gonna look we're going to move out of social sciences completely here we're gonna look at two examples, one from human medicine and we're going to look at the issue of artificial insemination and how we can use cross classified models with a data set to do with artificial insemination and actually we were looking at insemination in a slightly different frame here we're going to look at biology or ecology, probably more truthfully in our third and final example, and this is to do with small birds called great tits and how successful or not they are at nesting attempts using a data set from Wytham woods, which is near Oxford.

So let's move along. So these are extensions and as I've said the models we've looked at so far have really assumed, we have a normal response. The response variable and, in fact, actually the residuals from the model fit follow a normal distribution. In other words, the data has to be approximately continuous but as with the nested models that you'll be more familiar with and linear models more generally. it's possible to have other response types, you know data doesn't all come as continuous data or can be transformed into continuous data often data will be binary it might be a count or could be categorical and we're going to focus here, in these two examples on binary data. And, and we will have some more continuous data in the second example, but the Fife example what it has is two classifications, we've got primary school and secondary school and we cannot nest primary school within secondary school or secondary school with in primary school.

But of course there are many ways that we can classify data and we may end up with several candidate classifications, we may have more than two. So the extensions we're going to look at the artificial insemination example has 3 higher classifications and the great tit ecology example has 4 higher classifications so let's let's move on, then, to those examples let's look at the example 2 as the Fife example was example one, and this is data that originates with David Clayton and his collaborator Rene Echoard.

They analyze this in a paper in 1998 and David Clayton again looks at this example with our former colleague John Rasbash who we have mentioned, I think, earlier in passing, that they developed a new algorithm called the AIP, alternating imputation prediction algorithm for fitting cross classified models and they are in that paper from 1999, and this was the example that they they used so Davd Clayton used to work at the MRC biostatistics unit, and this is a medical example and what we're looking at here is clinics that help women who have trouble conceiving babies, and so what we have is a data set with lots of measurements of what we call ovulatory cycles, these are where women will have an attempt to become pregnant.

In each of these cycles and so we've got a nested data set here and, if we look at the top half of this unit diagram down here we have cycles nested within women, so we have a nice - this chunk here is a nice nested data set so women will generally keep having attempts -It won't be a kind of a random pattern usually what we'll find is that we have cycles, where if the woman doesn't get pregnant and then of course she will then get pregnant. And it's possible I guess that she would come back to the clinic and try again for another baby so we're going to take that data as 0-1 data so one means a baby is conceived during a given cycle so, there we are a nice nested model, but we have more information than that, because this is a clinic what we know is something about the sperm donors used for the cycle and so, so we have sperm donors at the bottom here and those sperm donors give donations. And the reality here is that each sperm donation can be split up, so it could be used more, for more than one cycle, so if we look at this example here we've got this first donor m1 and they've got 2 donations donation one and donation 2, donation one's been split up and it's and been used for women like woman one and woman 2 for cycle 2 and their second donation was again used for this first woman.

In cycle 2 and was used for woman 3 for cycle 4 and so what we have is, we have down in this part of the model we have lots of crossing and actually down here we have a nice nested structure again okay so donations are nested within donors and we can represent that in the classification diagram over here so we've got cycles at the bottom crossed between donations and women and donations themselves are nested in donors and that's nice it's nice to have this classification diagram to show people what's going on in the structure.

From a computational perspective all the computer software, for the MCMC algorithms at least, need to know is that there were these 3 higher classifications, and this relationship here is not really used in the algorithm I mean if the donor had been just here as an extra blob in the middle, it would have made no difference to the algorithm because we were assuming unique IDs so each donation has to have a unique ID each donor, each woman etc. Okay. that's that's what we've got.

In this example and we've only got - this is a very short example - we've got one more slide where we talk about the model and the results so here on the right hand side is our model. On the left hand side, so I don't know my left from my right, and here we can see the equations for this model, so we have y_i using this notation that was developed by myself and colleagues, the single subscript notation, so y_i stands for the ith cycle in this dataset and y_i. y_i takes value one if there was a successful pregnancy, Or I mean successful conception and zero if not and because it's binary data we use a Bernouilli or, if you like, Binomial with denominator one distribution with parameter pi_i. pi_i is the probability of the successful conception on this cycle so we're going to use a logistic regression type function here so we've got logit pi_i.

The logistic regression of that probability, and in this bit here X Beta we're going to put a whole load of predictor variables, things that we've measured for that cycle, they maybe things about the woman, they may be things about the donation, they could be about the time so and then we're going to say, well, which of the various classifications is the most important so here we've got 3 terms. We've got one random effect for the woman, the woman involved in that cycle, one for the donation that's been involved a specific donation of sperm and one for the donor who gave that to donation, and we assume normal distributions here, this is random effects for each of these terms.

We fitted this model using MCMC, in theory you could have done this in IGLS or the AIP algorithm that Clayton and Rasbash developed, we found it easier to use MCMC and we can see some results on this right hand side and what we can see then down the bottom here are the 3 sources of variation and we can see the relative magnitudes of these variabilities. So what that shows is actually the woman level variance is the biggest and then the donation and then the donor. So, in actual fact there's a lot more variation between women in terms of whether they become pregnant or not than there is for donations but even there some donations are more there is more variation there so there is more difference between two donations, than there is, between donors. So that makes sense to use different donations from the same donor as a result of that and, here are some explanatory variables that we've put in the model, this is in the X Beta bit.

So, so we have from firstly the intercept is very, very small and negative so that suggests quite negative to me i'm not sure. Whether these predictors have been centered or not, so I wouldn't try and describe that number at all, but what we can see, for example, is azoospermia this predictor here has a positive effect and that's sort of azoospermia is kind of meaning that we have poor quality sperm in a way. But actually the azoospermia is actually a condition so either you are azoospermic or not. And you might say, well, why would that be positive? well, this is actually a predictor associated with the partner of the woman not the donor or the donation, so what this is saying is why it's positive well if the partner is suffering from azoospermia then it's more likely to be him, rather than the woman involved, that is causing the difficulty in becoming pregnant, so therefore that such such women whose partners are azoospermic are more likely to become pregnant in this data. The quality of the semen used that's that's quite straightforward so that's a positive effect, so the higher quality, the better.

Both women and men's fertility reduces as age goes on so we have just got a dummy variable for that, with the woman's older or younger than 35 and there is a negative effect there. So the older women are having more difficulty. The count has a positive effect so if you have got a higher sperm count in this experiment this is seen as positive. The motility doesn't seem to make much difference here (the movement) to the conception that's not a significant effect.

Then we've got a couple of timing effects so if the woman is inseminated too early that's got quite a large negative effect and if it's too late, this is also a negative effect but not of the same magnitude. So there we are we've seen a cross classified example this top part where we're looking at the different variables that's very much like any other sort of logistic regression model but down here we can partition the variability so we can see, which are the sources of variation and it seems like women are actually more important than the donations that have been used in this case. So that's an interesting example let's move on, then, to our second example.

And this involves data that I personally analyzed. So this is the Wytham woods great tit data set, and this is a collaboration with the colleagues in Oxford at the Edward Grey Institute. Colleagues, Chris Perrins, Ben Sheldon and Richard Pettifor came on one of our workshops and brought this dataset along and we've analyzed it together, so what we have is a longitudinal study of great tits and basically this is quite time intensive data to collect so the team in Oxford every breeding season for the birds go to this wood not far from Oxford itself, Wytham Woods just on the outskirts and they put up nest boxes and they do all manner of collection over that period. So they measure the clutch sizes of any birds that is any nesting attempts that they find the date it was laid, so they are going, day after day round the nest boxes looking to see if any eggs have been laid. The mean nestling mass so when the nest actually comes to fruition, how heavy the nestlings are in the litter, clutch sorry not litter! wrong type of animal and then 3 sort of post event measures: was the nest successful and how we measure that is if any of the fledglings in other words the chicks that form this clutch. If there are any of those that survived to the next breeding season, then that would indicate a successful nest, we also have indicators of whether the male and the female were seen again breeding in future years.

So that's six responses, six measures that they've taken, as I say this is a very time intensive study. We've got 34 years of data. And over that period we've just got over 4000 nesting attempts so we've got 4 ways of classifying the data for every nest attempt, we know the female bird involved, the male bird involved, the nest box, because these are scattered all through the woods, involved. And the year of the nesting attempt. So year might be important for climatic reasons if it's a cold or hot year that might have an impact. The nest box again that might be territory, and you know, there might be better places in the wood than others, and then the genetics, of the birds to see what's going on there so we can summarize the data structure by looking at each of these sources in turn and we've got quite a few measures for each year they do vary from year to year, I think the data gets bigger as time goes on, but we have 34 years on average year is about 100 plus observations.

The nest boxes, they were quite a few of them about a 1000 for our 4000 attempts so we're only seeing on average about four. And for the male and female birds there are nearly 3000 of each of these for our 4000 nesting attempts so really we don't have a lot of information for each bird and when I saw this data, first, I thought there's no way we're going to be able to pull apart. You know male effects from female effects when the median number of observations you see for a bird is one, that would just be completely confounded but what we have you know is the mean is slightly above one and so some birds are going to give us more information because they're going to have several breeding attempts.

So another way diagrammatically to look at this there we go from 1964 - quite old data - to 1970 and we can see male and female birds and we can see male and female birds appear here we're the Female bird 1 has an attempt with male bird 1 then we come onto 1965 male bird 1 is still there and now it's female bird 5 and so on and so forth, so really we're seeing lots of singletons here pair 14 here birds 14 and 14 that's the only data point we have for them so we have no way of saying whether any effect was due to the male bird or the female but we've looked at these two down here but we've got some impressive birds, so for example, male bird 11 here mates twice with female bird 12, then the next year is mating again but this time with female 15 and, finally, with female eighteen. Actually, this an interesting pairing of 15 and 11 who mate together and then 15 has a new partner so we have all these different patterns here so we're going to try from the limited data that we've got to pull apart different effects.

So we're going to do univariate modelling and for each of our responses we're going to fit quite a straightforward model and we're not gonna fit any extra predictors. We are just going to fit our 4 levels or classifications. So we use that notation from my paper in 2001.

And so here we're going to have an overall mean for whatever our response is let us say clutch size for a minute and so clutch size has an overall clutch size Beta, it has an impact of male bird involved in the nesting attempt an impact of female, an impact of nest box, an impact of year and something to make everything add up a level one residual and we assume different distributions for each of these and we're interested in what the relative sizes of the different variances are because that tells how important the different classifications are.

We're going to use MCMC in MLwiN and we are going to use diffuse priors and run it for a long time. In actual fact we've got some binary variables there, the nest success ones where we run even longer, and we use diagnostics, to decide whether we tried it for long enough and we could - you can do some classical modeling if, if you remember, I mentioned some of the work in animal breeding and some of that stuff was put in Robin Thompson's work was put into the Genstat software package, so we can get results from that so let's look at- we're not doing model comparison - let's just look at the model with everything in it I haven't actually in this slide put any of the Genstat results, but if you were to look at the paper that looked at this, then you can see that as well, so here's the clutch size.

So these birds on average lay just under nine eggs and these are the relative different variation, variabilities so there's quite a lot of unexplained variation and that is not that surprising, with the sort of data that we have here but look in actual fact, it looks like it's really the female bird that drives how big a clutch size is and she lays the eggs, so that makes complete sense. So nearly 40% of the variation is due to the female bird. Virtually no variation is due to the male bird, there's a little bit of variation due to the nest box and some more for year so there are various hypotheses here, you know, that maybe the birds are adjusting how many eggs, they lay depending on whether the year is good or bad. So that's that's quite interesting and I can show that even with this data set with 4000 observations and 3000 of each type of bird we can pull out these effects.

So if we move on, then to the lay date. So these are days after the first of April. So on average these birds are laying right up at the end of April, beginning of May 29 days after April 1st. But there is quite a bit of variability here and this time around the biggest source of variability is the year. So climate makes a huge difference in terms of when the birds lay their eggs so a warmer or colder winter will make a difference. There is a difference again for females, maybe some females are more resistant than others and once again the males and the nest boxes don't have much of a say here.

Okay. So let's move onto the next measure which is nestling mass, so this is actually the average mass of the chicks at 10 days old. So not when they were laid, when they were like 10 days old and what we see for the first time here is that the averages in grams was, I guess, nearly 19. There's lots of unexplained variation over 60% of it we can't explain, but, for the first time the male birds have an impact, so here they're not quite as big an impact as the females, but that's kind of interesting right so 10 days, the birds share the collecting of food, so you would expect, maybe the male birds will now come into play, and year has an effect, obviously. Some years it's easy to collect food than others nest box again not much of an effect going on there.

Okay. So let's finally look at some of these success criteria, how successful are the birds, and we can see the value 0.01. So this is a logistic type model if we translate - its roughly zero - so remember definitely zero and the logistic regression is equivalent to 0.5 or 50% so 50% of nests are successful, in other words, one of the ringed nestling is captured in later years. What impacts that? well the biggest driver, there is the years. So some years the weather is terrible and that is going to kill off all the chicks, not all of them or the whole species disappears! but the number that will not survive would be bigger but there is, there are some genetics here so some male birds and some females are obviously having more success than others. And there's a little bit of effect for nest box only about 7% of the over-dispersion is due to nest box.

So we can similarly look at Male survival. And what would affect male survival? so once again year is the biggest driver but you know it's males that we're looking at so the particular male has an impact, the probability for back transform this -0.428 is about 0.33 and actually these numbers are all very small - we are looking at percentages here, but there isn't much over dispersion and they actually here, we are seeing them observed, they have been observed breeding in later years, so the real probability is probably a little bit higher - now most birds that survived to next year will be part of the breeding mix. but that's what we're seeing this and we can see a similar picture if I clear these drawings and move on to the last slide here on data here's the females and actually the females, interestingly again, there's a slightly higher probability here. Its 0.381 of them being observed later again there's not a lot of over dispersion, but this time around it's actually the nest boxes that they're using and the year that are having having an impact now so maybe there's some geography thing going on, but we don't over interpret these because these are over dispersions that are quite small.

So if I clear those drawings so that's anothe,r that's a second example of using cross classified models in practice so that's the end of the lecturing, and so what have we covered? so in this third lecture we've shown how we can extend a cross classified model to 2,3,4 different higher classifications and also to different response types. We looked at some examples here again of continuous variables, but also of binary variables, and we can have poisson models, we didn't have any examples of those here, but for counts, or even ordered multinomial or unordered multinomial models. We saw how we could use the notation and the classification diagrams for these extended models.

We saw two examples, the first example we were looking at how different factors within an artificial insemination data set and we saw that women were more important than the donor and the donation in terms of whether the attempt would be successful and we've seen in our second example lots of predictors related to nesting attempts and we've shown how we can partition and get biologically reasonable results you know where female birds are the driver on things like laying eggs that the females have more control over but actually when things like nestling mass where both male and female have a role to play in feeding the chicks male birds actually have an effect so hopefully that's the third example we've also looked at the earlier.

Fife data example in Scotland, educational example, hopefully now from these three lectures you have got a good grounding of the types of models, you might use cross classified models for what types of applications we're going to finish off after this then we'll also put up on the website a cross classified practical.

So here, you can watch, you can try the practical yourself, or you can watch the walkthrough video and it would cover the fight and Scotland example at George covered in the second of these talks, so thank you very much for listening and I hope you'll join us again when we look at multiple membership models, thank you.