Multilevel models for clustered dichotomous outcomes, part 1—2010 MCH EPI pre-conference training workshop

DON HEDEKER: Okay, so what Subash has worked through with you this morning and the early afternoon is clustered data where we have a continuous outcome. What we're going to talk about today, this afternoon is the same kind of setting, but where the dichotomist- the outcomes is dichotomist, so yes, no outcome. Obviously in **** you get a lot of dichotomist outcomes, so the basic structure of the analysis is similar, but essentially the left-hand side of the equation has to be changed because we no longer are modeling a continuous normal outcome, so we're going to be speaking in terms of a logistic regression kind of multilevel model. That is the basic idea.

Now some of you I heard some discussion about getting some of the datasets and the SAS code and all that. I have websites and here is one of them where much of this material can also be downloaded from as well. You can get the PDF files as well from the website there, so if after you leave if you want to carry out exactly the same analyses, the syntax you can get it all right there. I also have a site where I have some publications listed and this one of them that I worked on recently that talks about the material that we'll be going over this afternoon, generalized linear mixed models in the *Encyclopedia of Statistics and Behavioral Science.*

Okay, so we'll begin the dichotomist outcomes and therefore what we'll be describing is an extension of your ordinary logistic regression to allow for random effects and therefore we'll be describing a mixed-effects regression. Then if we have some time we'll also describe models for ordinal outcomes and those would then be mixed-effects ordinal logistic regression models and the most popular form there is what is called the proportional odds model where the regressors have the same effect on the ordinal outcome. Of course, that can be extended and it has been, into what is called partial or non-proportional odds model and then interestingly enough there is connections between these models and survival analysis when survival is measured at discrete or group time points. So let's say you're measuring some kind of outcome across several follow-ups and you don't know exactly when the event happened. You only know it happened here or between here and here or between here and here. There is a finite number of intervals in which it can happen. That is what is called group time survival data and logistic regression models can be used to estimate such survival analysis models and so if you have a clustered design you would then have a mixed-effect survival analysis, so that is another connection to these models and worked on some chapters on that as well if you're inclined, interested.

So we're start with the dichotomist outcome first. So just to review, before we get to the multilevel structure let's review a little bit about logistic regression modeling. So here on the left-hand side of the equation we have the logit, the log odds of the response. We typically code the binary response zero or one, but realize of course that is completely arbitrary. We could code it one or two, minus one, one, whatever. It's convention to code it zero, one and to include in the logit model in the numerator the probability of the one response and in the denominator the probability of the zero response, one minus that, so what we have in brackets there is the odds of the response and the log of that is called the logit or log odds of the response and you can think of the logit function as a function that links the probabilities to the regressors because on the right-hand side of the equation we have our familiar X beta, our multiple regression model and in X we can include all the usual suspects. We can have continuous regressors. We can have grouping variables. We can have indicative variables, interactions, polynomials. Anything that you learn about an ordinary multiple regression can be applied here as well in X and the betas of course are their coefficients.

So it's convenient to write the model in this way, the logit is equal to X beta, but you can also write the model in terms of the probability, so using some algebra you can convert this equation to this equation that shows that the probability that Y is equal to one is equal to this nonlinear function. This is what is called the logistic response function. If beta is positive essentially looks like an S-shaped curve, which as X goes up the probability goes up, but of course it has a lower **** zero and a higher **** one because probability lives between zero and one. So another way, mathematically a way to think about it is like in statistics, in modeling we like to have X beta, but our outcome in a binary is a probability. It's a bounded sort of space. Is there some mathematical function that can convert that bounded space to an unbounded situation? That is what the logit function actually does because what it takes is something that goes between zero and one and instead it goes from negative infinity to positive infinity, so it takes a bounded statistic and makes it unbounded and therefore on the right-hand side we have X beta. So you might ask yourself why can't you just put Y is equal to X beta taking the binary outcome. Why would you not want to do that? Because X beta wouldn't know that it could only be zero or one. You could easily get negative estimates, estimates that are greater than one, so using just Y equals X beta doesn't make any sense for a binary outcome, but using the log odds then makes it work basically.

So one thing to realize too when you look at these two equations is that the model is a linear model for the logits, meaning here we have X beta, the linear model, linear in terms of parameter, but here we do not have a linear model. It's a nonlinear model. That is why sometimes logistic regression is called is a called a nonlinear regression model because essentially this form is not going to look like a line because again, the reason it doesn't look like a line is because probability goes between zero and one, so you can't expect a line is going to work well in that scenario. So logits take on any values between negative and positive infinity. Probabilities only take on values between zero and one. So here is what the logistic response function looks like for a- if you have

a positive coefficient for and X variable, an S-shaped curved. That S **** down at zero and S **** up at one. Taking the logarithm of the logit we get the straight line.

Now it's very important. Probability is a scale I think most people understand pretty well. Logit less so perhaps, so one thing to think about is well is there some sort of bridge that can help us understand a logit transformation? Well think about this. Probability of a half, what does that correspond to in terms of odds, probability of a half? It corresponds to odds of what? You have probability of a half over one minus that. It's a half. That is odds of one. What is the logarithm of one? Zero. So the probability of a half is like the anchor point. The probability of a half corresponds again to odds of one, corresponds to logit of zero. What that means is if you get any positive logit value that implies probability greater than a half and correspondingly if you get any negative logit value that corresponds to probably less than a half. So that is the first things to realize with the logit scale. Negative values of logits implies probabilities less than a half. Positive values of logit implies probabilities greater than a half. Zero logit value corresponds to probability of a half.

You can also write the model in terms of the odds. It's not so common to do that, but what this helps us understand is why everybody in this room when you get a beta in a logistic regression what do you automatically do? You take the exponential of it, right, to get the odds ratio. So why do you do that? Well if you write the model here in terms of the odds, so notice what I've done I've taken the log, the **** log over here, which is the exponential, so I now have the odds equal the exponential of X beta, so exponentiating the beta gives us a change in the odds of the Y per unit change of X and this is a natural interpretation that really makes logistic regression so much more popular than other binary regression models. For example, probit regression, that is another alternative methodology for binary regression, but probit doesn't have that nice interpretational quality like we get from the logistic regression where we can just exponentiate the betas and we got a pretty understandable meaning of what that represents. So here if we have beta equals zero again. That would indicate no effect of the odds. Beta greater than zero would increase the odds with increasing X and a negative beta would mean that the odds diminish with X increasing.

Now this is some part of logistic regression you might not have been exposed to, but I like to make use of what is called the threshold concept to help me better understand the meaning of the parameters, the betas in the model. Let's suppose for the moment we've got this binary response here YI and underlying it is the continuous response strength Y in the italicized version. It's unobservable. We don't see it. That is why it's latent. We don't see it at all. What we do observe is the manifest variable YI, but these two are related via threshold concept and the idea is this. If a person's response strength, which we don't see, is above some unknown threshold then we get a positive response. Alternatively if the response strength is below the threshold we get a zero response or a negative response and what this shows is the two distributions, possible distributions for this underlying response strength, the normal and the logistic and in their standard form they're centered around zero and what is the variance of a standard normal, it's one, right. What is the variance of a standard logistic? Well you can see

right here it's going to be wider than the normal. It's more diffuse. They're both centered around zero, but they're more diffuse. The variance of a standard logistic is Pi squared divided by three. Now whenever I say this I feel like a magician pulling the rabbit out of the hat. Where did that come from? Well you could go into the derivations of that, but you're probably not so interested. Let's say it came from **** book on categorical data analysis and there you could find the derivation of this, but the variance you can appreciate of the logistic is wider than the normal just by this plot. These are both the standard versions of these two distributions normal and logistic.

Now why I like this threshold concept is this. Have any of you ever run a logistic regression and a probit regression on the same dataset? You should it sometimes. It's a lot of fun. What you'll find out is that the estimates from the logistic are numerically larger in absolute value than the normal and likewise, their standard errors are a lot- there is a scale difference between the two. Why, this plot helps us to understand that because the scale of the logistic is wider than the normal. So in fact, we can write this model logistic- and here shows a little slide about this threshold concept that tries to make it so that it's not so crazy. So after today somebody might ask you how is your day? What is your satisfaction level today? Well obviously satisfaction is a continuous kind of variable, a matter of degree where you're not going to say on a scale of 0 to 100 I had a 98.7. You're not going to say that. You're probably going to say I was stuck in a statistics conference. It was a negative 2,000 or something like that. No, you're likely to say I had a great day or I had a day. You'll emit a binary response, maybe an ordinal one, but let's say binary when there is this underlying continuous variable there.

So here is the model written in terms of the underlying continuous variable, Y equals X beta plus error and again, if we assume that in this model that the error comes from a standard normal we have a probit regression. If we assume it comes from a logistic we have logistic regression, so when you run PROC Logistic or whatever logistic program you're assuming this basically. You may not realize this, but this is what you're assuming and in this version of the model again, I make reference to this because it helps us to understand the scaling of the betas. What do the betas reflect? Betas are scale translators. They tell us as X changes by one unit how much does Y change. Now notice between here and here the scale of Y is different. In the probit the scale of Y variance is one. In logistic it's Pi squared divided by three or approximately three. Just think about that. So what that means is that the beta estimates from logistic regression are larger in absolute value from probit by approximately this factor. Here I'm just taking the standard deviation of the logistic variance, approximately a factor of 1.8.

So you know many times when I was beginning I would run logistic, I would run probit I would see the scale difference like why would- why? I never understood this really. You make reference to this latent variable that underlies it you can see why you get a scale difference. You've changed the scale of the dependent variable flipping from one to the other. That is why the betas change. You might say if that is the case I'm going to run probit logistic all the time. The betas are going to be bigger, right, so I'll get more significant results, but the standard error are effected the same way, so in fact, the ratio

of the estimates of standard error between probe and logistic they're virtually the same. It's not an equality, but they're approximately the same. They're very, very similar. It would take a very large dataset for you to see much difference between probit and logistic per se in terms of inference.

So I want to make this point that this underlying linked variable is a useful way of thinking of this problem. It's not an essential assumption of the model. What it helps us to understand at this point is why the betas are on a different scale as we go from probit to logistics. Now why do I make reference to this? The betas are going to be on a different scale with or without random effects as we bring those in the model, so we're going to use this concept to help us understand the scaling of the betas.

Now I want to say one other thing before we move on. In what you've covered up to this point with a continuous Y outcome you don't see this. Why don't you see this? In continuous Y models you estimate the error variance and so the betas are the same in terms of numerical scale with or without random effects, with or without the different kinds of X variables in. It doesn't change, but in these binary regression models here notice this is not estimated error variance. It's fixed. The consequence of that is that the betas are on different numerical scales here when we go from probit to logistic likewise when we go from models with and without random effects. So again, if I write the model in the ordinary logit manner I don't necessarily appreciate the scale difference, but when I write it in terms of the latent continuous response it helps me understand it to a better degree and that is why I make reference to this. Although again, if you're running this with software you wouldn't necessarily need to know this per se it helps us to understand the basic of the underlying model.

So now let's introduce random effects. Let's say we have cluster designs. You have patients nested within hospitals, patients nested within providers, something of that nature. So now we got a model say with P covariates from the response. YIJ, subject I and cluster I. Sorry, subject J and cluster I. Now a very important number up there is that J goes from 1 to NI. Everybody hates notation, right. I always felt the notation is your friend. It's your friend and they don't believe me, nobody. Subash believes me I think. You've heard this spiel so many times. NI, it's not N. It's NI. Why is that important? What would it mean if it would just be small N there? It would say that every cluster has the same number of observations; every hospital has the same number of patients, every provider same number of patients, something like this. Did you ever see that with real data? No. With textbook data, yes. Real data, no, not so much. So it's important that our model can accommodate that incomplete, that differing data. Tomorrow we're going to talk about longitudinal. What that is going to imply, the fact that we have NI is there is that subjects have different numbers of time points because in longitudinal data analysis that is often the case. People they may have been- there may have been an intent to measure them at five time points let's say, but not everybody is measured at all five, something like that. In the clustered scenario what it means is not every cluster has the same number of subjects within it. Again, that is pretty common with real data, so we're going to allow that in these models and that was allowed also this morning in the models Subash was describing.

So here notice we've got essentially the same model we had before, a logistic regression model, but we've added in this one parameter here, so first thing to note is these mixed or multilevel models that we're describing today and tomorrow they're really just augmented multiple regression models or in this case augmented logistic regression models. What this means is again everything you've learned about ordinary logistic regression really applies here as well, so X your covariance can be just like you have in your ordinary logistic regression models. All we're doing is to add on one parameter at this point to this model that you used all the time. What is that effect? That is the effect of the cluster on the person's probability of a response, so in other words some clusters have a greater chance of a probability of event. Others clusters have lesser. There is a cluster, a fact on the individual. The people within a cluster are correlated to some degree. They're not just like a random sub sample. There is something systematic about the people in the cluster. People seeing a particular doctor might do better than the people seeing another doctor, something of that nature. This is the parameter that is going to capture that.

Now in this kind of model we're thinking of the clusters as coming from a population of clusters. That is why technically this is called a mixed model. The betas are fixed regressors. They're fixed. What does it mean in statistics to be fixed? There is no distribution. There is just a gender of fact that we're trying to estimate. There is a difference between treatment A and treatment B that we're trying to estimate. The cluster effects alternatively we're going to have how many of these in our sample? We'll have large N of them, but we feel that the clusters we have in our sample come from some population of clusters that we want to generalize to, so the hospitals we have in our particular study are representative of some larger population of hospitals. Maybe they were randomly selected. Maybe they weren't randomly selected, but we conceptualize that they have come from a population that we want to say something about, so therefore, the cluster effects have a distribution in the population and here is what we're assuming about that distribution. Like I said it's normal. The clusters are independent of each other. That is what the I is about and D is for Don, right. No, distribution. D is already gone. Got to come up with better things than that. So it's normally independently distributed with mean zero, so they're deviations from X beta and variance Sigmund squared epsilon, so Sigmund squared epsilon is what describes how are the different clusters different from each other. Let's say cluster has no affect on the subject's data, so the subject has no affect. The hospital has no affect on the patients within the hospital in terms of their probability of some disease or something of this nature. What that statement would- What would that mean? It would say that all of these epsilons are zero. They have no effect. What does that imply about the variance? It goes to zero.

Now let's say alternatively hospitals have a big affect. Some hospitals are really good ones, some are really bad, probably what's more like reality I suppose. These will be different from each other. This variance will get large, so the population variance of the cluster effect tells us basically how much effect the clusters have on the individual data. If it's a big number then that reflects a bigger affect. If it's a small number, a smaller

affect. Now what you learned about this morning is that a convenient way of expressing the cluster effect is in terms of its ICC, which is basically the cluster variance divided by the sum of the cluster and error variance or what this is sometimes called the level two variance and the error variance is called level one error variance. Why do we express it as ICC? The metric of this variance is arbitrary. It kind of depends on your dependent variable. However, an ICC is kind of like an R squared measure. Think about when you run ordinary regression modeling. What is often reported as R squared? What percentage of the variance are you explaining with your regression model? And it's a common metric. It goes from zero to one, so no matter what you're modeling it's something that everybody can understand. It's like well R squared if you're not explaining anything it's like around zero. If you're explaining a lot it's around one, likewise the ICC. It goes from zero to one. It's a proportion of variation that is at the cluster level, so regardless of the dependent variable you're modeling everybody can kind of understand ICC because it's a common metric.

Now a clustered data as you saw this morning with the example with schools and with classrooms the ICC tends to be around 5%, maybe 10%, in that range. Not that big, but realize if a school is having that much affect on the students within that is actually kind of a big affect because a school is a pretty diffuse, it's a large entity. When we get to longitudinal tomorrow obviously the ICCs will be much greater because people tend to react similarly across time more so than people react similarly within a classroom and so that is the first basic rule is that for clustered data the ICCs tend to be rather small as compared to longitudinal where you're modeling the same person across time.

So some characteristics of these random effects, again- Yes.

[Comment]

The question is all about the sample size requirements in terms of the clusters as well as perhaps the individuals within clusters, things of that nature. Generally with regard to clusters some- Sniders and Boskers for example give these rules of thumb. If you have 10 or less you don't have a great ability to estimate this cluster variance because you basically have 10 observations let's say, 10 or less, so let's say you've got states. Let's say you've got five states. Would you want to treat state as a cluster, as a level of the analysis? Probably not, you don't have a lot of information between states. Instead what you can do in that case is maybe not conceptually ideal, but pragmatic, is just to include indicator variables for the five states. Put four dummy variables in to account for state-to-state differences. What you won't be doing is to say that state effects come from a population of states. That is the essential difference between treating it as dummy variables, fixed effects versus random effects. If you just put in indicator variables to account for the differences between say the five states what you're saying is you're only interested in these five states. There is no population of states from which these come from, from which you're generalizing to.

Now if you have five states you might still think this is a sample that I'd like to say something more than just about these five states. The problem is you don't have a lot

of data by which to do that, so conceptually you might want to do that, in other words, treat it like a random effect because you think these states they sort of represent their US states. I don't have a lot of them admittedly, but I would like to do that. The challenge is again you don't have much information by which to do that. If you have over 15 people generally say or 20 or so then typically no problem treating it as a random effect, treating it as a level in your analysis.

[Comment]

Yeah, it is actually. Yeah, I'll put this away unless you like- People like this. My daughter loves that sound because she is that's cool because it does the same thing with the radio actually at home when I'm charging it really close to a radio. I don't know what this says about her musical taste or lack thereof, but okay. We'll remove it for today. Yes.

[Comment]

Yeah, there the requirement for sample size is not so important in a sense because typically you've got a fair number of clusters. You still have to think about well let's say you only got 2 or 3 subjects per cluster and you've got say 30 clusters. I don't know, something like that. You then can entertain only so many subject level predictors because you don't have some many subjects per se, but strictly speaking other than that there are no inherent problems. Usually the sample size problem is more about the number of clusters I think because that is where you're more often limited in terms of numbers. Realize this too. You might say well I've got 1 cluster that has got 30 subjects. I have another cluster that has got two. Should I put those in the analysis? Well realize this. They give you very little information. It doesn't matter really if you put them in our out per se. It's not like putting them in is going to greatly change things. They're so small, so in that way again when we get to longitudinal people often have this idea well if a person is only measured at one time point should they be in the analysis. There is no inherent harm in putting them in because they provide so little information anyway and if you're going with an intent to treat sort of point of view you would put them in. Again it's not like they're going to change much because there is not much information there.

So N is typically the thing that you got to think about a little bit and admittedly when you're sort of in the 10 to 15 range there it's a little bit of a grey area. Should you treat it as a random effect? Should you treat it as a fixed effect and just have indicator variables and depending on discipline the answer could vary. For example, in econ they're more likely to do this as fixed effects. For whatever reason they don't really prefer random effect models as much.

One thing I should say is that people have done simulation studies in this regard and it is true that even like with 15 you're not going to estimate this variance all that well, the cluster variance, so if you're really interested in that, how different are clusters from each other, you need to get a fair number of clusters to get a good fix on this. So you

can treat it as multilevel at say with 15. You're just not going to know this with a great deal of precision. Yes, Subash.

[Comment]

True, one thing to realize is once you go from continuous normal to binary or accounts or ordinary or anything like that computationally much more challenging, so what that implies I mean just think about this. You've got a continuous normal outcome. There is a lot of variation, a lot of information there. With a binary you've got two values. You're trying to model variation and there is not a whole lot there, so computationally there is more things that could potentially go wrong in these scenarios. Now we try to get over those things as best we can, but it is true that continuous normal models are estimate much easier than logistic or **** or etcetera.

So the random effects are what separate the model from the usual fixed effect. We have 10 random- sorry, 1 I to- 1 N random effects. It tells us the impact of a cluster on the logit subject within the cluster. It represents the degree of subject clustering. It's the same for each member of the cluster, but changes as you go across clusters. If the cluster effects are all zero then the cluster has no effect. If the effect is zero for all clusters then the cluster structure has no impact. What you're saying is that the cluster variance equals zero. In that case there is really no need for a multilevel approach. Ordinary logistic regression is okay. If subject clustering has a strong or even a moderate effect then these will not be zero and the variance is going to increase to zero.

Now something that came up and I don't mean to say differently than what Subash told you this morning, but I know a little bit more details about that particular study. In that three-level study their schools were randomized to condition and the intervention was delivered in the classrooms, so from a design point of view it seems like the school should be representative as a level even though as Subash pointed out it's not really adding much significance to the model and clearly a two-level model was adequate for that dataset, but from a design point of view because the school was a unit of randomization and these days we can estimate that three-level model quite easily relative to when that article was published as well in the mid 90s that would be the better approach because even if it's significant or not to put it in as a random effect, put it in as a cluster because of the fact that in the design of the study schools were randomized and the design was administered in the classroom, so therefore logically regardless of the testing the three-level structure is really- seemed to be the most appropriate one.

So here is now the model written in terms of the latent response. So again, we don't see this YIJ, but we can write the model in terms of this. Now notice this. If we have a random intercept model on **** there is two sources of variation. The error variance, which his assumed to either be a variance one if it's probit or Pi squared divided by three if we're choosing logistics and the cluster variance. What this is going to imply is that the betas again are on a different numerical scale with or without the random

effects because this, if these are non zero then the variance of Y is going to be bigger if we had these in relative to if we don't have them in. So again, that is why I make reference to this latent continuous way of writing the model. It helps us to understand when we run an ordinary fixed-effects model and we see these betas and then we run a mixed model it seems like the betas are pumped up. Why is that? It's because of this because now there is two sources of variants influencing Y and in the model without the random effectors on the error of variance.

So again, this underlying latent variable nonessential assumption of the model is useful for obtaining the ICC. Here is the ICC. In the dichotomist representation we have the cluster variance divided by the sum of the cluster and error variance and the error variance here is not an estimated quantity. For logistic it's assumed to be Pi squared divided by three and this also related to something you see in the sampling literature called the design effect, which has in the numerator what is the sampling under cluster sampling. In the cluster sampling you have cluster variance plus individual variance relative to what the variance would be under simple random sampling in which there is only individual variance. So these two ideas are related, expressed somewhat differently depending on the literature in the sampling literature you have this relationship, the ICC that. So the cluster effect ratio of the actual variance- sorry, design effect, ratio of the actual variance to the variance that would be obtained by simple random sampling.

So again, to understand the scale of the betas the betas from logistic regression are larger than from probit regression by approximately this factor of 1.8. Why, because the variance of Y is equal to Pi squared divided by three and this is the underlying Y in logistic. It's equal to one in probit. So all I'm doing here is basically taking the ratio of the standard deviation, so I'm equating them on that scale. Likewise, if we compare a random intercept or mixed model to a model without the random effect, a fixed-effects model. The beta estimates from the mixed model are going to be larger in absolute value from the fixed-effects model by approximately the square of the design effect. That is the variance under the mixed model, the random intercept model relative to the variance under a fixed-effects model.

So what this is going to suggest is that when we run a model, a logistic regression model we're going to get these betas, then when we include the cluster effect numerically they're going to be larger in absolute value and it's because of this. Obviously it depends on the size of the cluster variance, so this is one thing, sort of a subtle thing. If you run clustered as we're going to run first you may not realize this because in clustered remember the clustered variance is typically not going to be that big. It's going to be non zero, but it's not going to be like huge, so you might not appreciate that the betas are on a different scale. They won't look that different. When you switch to longitudinal they'll look very different and really that is where it first hit me. I was running these kinds of models in the 90s and didn't realize the scale difference until I got to the longitudinal part. Then I ran them and I was like wow. Again, these betas, they look like they're pumped up. They're on steroids or something. What is going on? That is what is going on.

Now it's convenient and nice to write the model in terms of a multilevel or hierarchical linear model. Here within a binary response we can do that as well. Here we write in the within clusters model we have the response and here I've written it either as the logit or the latent response and on the right-hand side I would include any subject level covariates, so let's say we got the person's gender, so that has and I and J because it's the subject J within the cluster I, so at level one we'll include any level one covariates and at level two those coefficients can come down to be modeled in terms of group level variables, cluster level variables and random effects.

Before the break there was this question with the school example about treating pretests as a random effect because pretest was like this. It was a subject level variable and it's true. You could then have that as a random effect. What that would represent is a model in which you think the relationship between pretest to post test is different from school to school to school. You allow that coefficient to vary between the different schools. Now if you were just wanting to account for the structure you typically don't go down that avenue, but if you have interest in whether or not the pretest, post test relationship varies across the clusters that is when you would entertain that kind of more advance model. Here we're just accounting for the clustering of the data by having the cluster random intercept, but admittedly it's true. You could have these models with multiple random effects and as Subash mentioned in longitudinal typically we do, not so much in clustered, though you could do that.

So we can write the model in the multilevel manner. Again, when I teach it in biostat I often do this. In biostat you're not so likely to see it written as multilevel. You see it more in social science literature and admittedly the first time I did this was I taught a class in sociology in multilevel and when I would see these multilevel models I'm like, "What is going on?" "There is one model." "What are all these models?" I was reluctant to do it, but I had to do it because I was teaching in that subject arena and the more I did it the more I realized it really helped me understand the model parameters much, much better. It helped me to interpret the results better. So sure, you can run these models without doing this. You can publish without doing this, but I would say if you really want to understand the parameters of the model not a bad idea to write it in this HLM multilevel kind of way because can really you know like here in this model let's say this is gender coded zero, one. Let's say it's coded zero for males, one for females, so this is the male intercept. Now we're saying that the male intercept is influenced by the group that they're in. The female difference from males is also influenced by the group that they're in, so it helps us to understand that. It also helps us to understand again this idea that you could have any level one variable could be random at level two. If you just go the software you wouldn't necessarily know this, but writing it this way you say wait a minute, can't the gender difference, can't that be random. Yeah, it could be the case that as you go from hospital to hospital gender effect is very different. I mean that could very well be. You write it this way you see that is possible with this kind of modeling.

So we'll see some examples of this. Yes.

[Comment]

Right, good question, so let's say we run this- Sorry, I don't know why this is-

[Comment]

I'm too loud. Is that better? Hopefully, we don't want to hear boom, boom, boom. I play in a rock band. It's a little much for me even. That's saying a lot. My ears are shot. I'm used to that kind of feedback as opposed to the good like the kind you've just raised. You're exactly right. Let's say we don't have a group in here for the moment and you just have the random intercept, then this represents how are hospitals different from each other. Now you want to try to explain some of that variation by bringing in **** group here, say treatment group or whatever hospital variables. Essentially what you're- your aim there is to make these smaller and smaller to explain some of the reasons that hospitals are different from each other with these hospital level variables and again, writing it this way gives you this notion because what happens in a ordinary regression model? You can think of this and in fact in the multilevel literature these are called the level two residuals. What happens to these residuals if we add in variables here that explain variation? These shrink. That is why I say your aim is kind of like driving them to zero as much as you can in a way to explain some reasons why they're different from each other and that is sometimes why people what they'll do is they'll first run the model without any regressors to get an idea of well how much cluster variance do we have, now let's try to explain that and those variables would go right in over here like I have here. Yes.

[Comment]

Correct, right, so then what you want to do is you want to take a look at the estimates of these because these will be the estimate of each cluster effect on the outcome, so you'll see which hospitals have higher or lower probability of the outcome based on these. Now admittedly in standard running of the problem you don't automatically get these, but I'll show you some ways of getting those out if you're interested in those. What you usually get is just this variance, the variance of these, but you don't automatically from the software get the actual cluster effects, but it sounds like you're interested in those to know which clusters are good ones, which ones are bad one. You're exactly right. From standard running of this as you have probably seen up to this point you'll get the variance of the cluster effect and that will tell you that the clusters are different from each other, but it won't identify which are better, which are worse.

[Comment]

Right, so it sounds like what you're describing is where you're accounting for the clustering, but you don't have any cluster level variables per se. Right and but you still want to find out which clusters are better or worse. That information is gotten from the estimates of the cluster effect and again you don't automatically get those out from

standard run. There is some additional options that you have to specify to get those things out. In fact, in the British literature they call these I think school table analyses because they're real interested in ranking the different schools and they use these estimate from the mixed multilevel model. Over in England they always call these multilevel models because Harvey Goldstein and his group developed this with multilevel software and so everything- Here in the States it's HLM. There it's multilevel; but anyway, they base it all on those estimates.

So I want to go to this example that we've seen again. Now admittedly this example is maybe not an ideal one for the purpose of the subject matter that you're interested in. I understand that and I apologize about this. This is not a maternal and child health outcome per se, but this example is a very good one for showing the effects of the model on the conclusions of the study, so what sometimes makes a really good teaching analysis in statistics or dataset it doesn't always have the optimal properties in terms of the substance of area per se, but it is a really good teaching dataset, so that is why I'm using it here and so again as you've heard this morning this is from more of a prevention science kind of study in which they're trying to prevent and stop kids from smoking. That is basically the outcome that they're interested in, in this particular. Now you might think of it in your own work. This might be trying to get pregnant women to quit smoking or to stop smoking or something like that, so you could translate it in your minds to something that may be more pertinent to your own work. It was a study conducted by Brian Flay. I often work with Robin **** that is why I started to say Robin ****, but this was Brian Flay. A three-level structure, we had seventh graders within classrooms within schools, widely unbalanced, 1 to 13 classes per school, 2 to 28 students per class. The outcome that we're going to look at is the knowledge of the effects of tobacco use. Now you've looked at it so far as a continuous response. What we're going to do this afternoon is treat it binary. Again in general I would not do that because you're losing information by doing that, but to make connection to a dataset we have some familiarity with this is why I'm using this example, so it's not optimal in the sense of like again by dichotomizing the outcome we're losing some information, but it makes connection with something that we already know a little bit about.

Now the students were tested at pre and post intervention and again the schools were randomized to one of four arms. This is essentially a two by two crossing of social resistance classroom curriculum, a media television intervention, the two combined and then no treatment control. Some people always ask me well what is this classroom curriculum all about, what is the TV intervention all about. I'm the dumb statistician. To me it's just X, but let me give you the dummies guide to these variables. As I understand it what the social resistance classroom curriculum meant is somebody walked in the classroom and told the kids all the bad things about smoking and if you're smoking stop, if you haven't started don't start. The TV was somebody came in with a TV, plugged it in and they played the same thing over the TV. It actually wasn't like that. They went home and watched TV, but it sort of makes a better joke if you say somebody walked in and- The combined they got both arms and no treatment they didn't get either of these interventions. Now what they're interested in, in this study is that typically if you have a person come into the classroom and talk to the kids that

usually does pretty good in terms of interesting their knowledge and their understanding, but you can imagine that if the television is successful that could be very cost effective because you don't have to have somebody coming in to every single classroom. You just produce that lecture once. It's like distance learning I guess in some sort of way and then boom, boom. You give out that lecture again and again and again. So they're interested in contrasting these two forums of deliver of the same basic message.

So the outcome of interest we're going to look at is the intervention effect on this knowledge score. The challenges, okay, the outcome variable is really a number of correct of seven items, so the kids answered seven items and they were scored whether they got it correct or not. We want to control for the multilevel structure and we have potential explanatory variables at different levels. I put those as challenges. They're really not challenges. They're easy to handle. So you want to always set your aims modest, so you can achieve them. So we can handle those.

What we're going to look at this afternoon is looking how many of the kids got three or more out of seven correct, so we're not talking about great knowledge levels here. We're talking about kind of like almost like better than chance or something like that. They're not doing all that great, 3 out of 7 and this show the 4 arms, approximately 400 kids per arm and the proportion of kids post test that got this response correct or more so correct, so in the control arm about 40% got them right. In the kids that got the TV, but not the classroom curriculum almost half, so it does seem like there is some good affect of the TV, at least within this arm. Them among the kids that got the classroom curriculum they didn't get the TV, notice about two-thirds almost. If they got both interestingly they didn't do the best, so that is why this data is kind of interesting. There is a bit of an interaction here. The TV seems to work pretty well here if they didn't get the CC, but combined not so well. The classroom curriculum has a pretty big affect. I mean this difference versus difference. This difference versus this difference is a pretty large affect, so again this is a dataset is kind of interesting because no matter what you do the classroom curriculum affect usually comes out no matter what you do because it's a strong affect. That's why if you really have strong data you don't need me or Subash. You need us. You can do simple analyses. With strong data things come out. Unfortunately, most of our data not so strong, more subtle, more interesting and so there the modeling does make a difference.

So this is a proportion. It's a metric we can certainly understand. Now in the next row I've represented those as odds, simply taking the proportion and dividing by one minus that, so again, odds that are less than one reflect proportions less than a half. Odds greater than one reflect proportions greater than a half, so there is only these two that bounce up over a half and these two are still over one and under one because of the proportions we see there. Correspondingly log odds or logits here we have negative values because proportions here are below a half. Here we have positive values because the proportions are above a half, so again, proportions metrically we can all understand. Odds, if you go to the racetrack you probably understand that too and

probably a lot of you understand odds pretty well. The log odds maybe not quite so well understood, but it's just the logarithm of the odds.

So let's fit a model here first that has, and this will address its point here, on level one we've got nothing, so we're saying that logit of subject J within cluster I this is just then basically the intercept for a given cluster or really you can think of the log odds of a positive response for the cluster and now what you're saying is that the cluster effect can be attributable to either the classroom curriculum whether they got it or not, whether the kids in that cluster got the TV or possibly the interaction, but these things are not going to account for everything, attributable cluster effect. There is still this random component at the end.
So this is a model in which we have no level one variables, but we have several level two variables. Now all of these level two variables are very simple variables. They are just indicator variables zero or one. So in this model what does beta not represent? It's the log odds when all of X equals zero. That is what it is equal. Well what does that correspond to in more English language? It's for kids that didn't get the classroom curriculum, did not get the TV. That is what beta not represents, the log odds for the kids in the control group. Beta one represents a difference between the control group log odds and the kids that got classroom curriculum, but not TV. Beta two represents difference in log odds between the control group and the people that got- kids got TV, but not classroom curriculum and beta three is basically saying how the classroom curriculum effect is different between kids who did or did not get TV or the TV effect is different between those kids that got or did not get the classroom curriculum, so the betas have the interpretation on the log odds scale. Exponentiating them gives you the odds ratio scale, so beta not logit for the control group. Beta one is the logit difference between- for the- that is for the classroom curriculum effect for TV equals no and why put this latter part? What I've done is to rewrite the level two model slightly aggregating the parameters around the classroom effect and when you do that you can see that aha, beta one represents the classroom curriculum effect when TV equals zero and beta three, the interaction is basically saying how does a classroom curriculum effect as you move from no TV to TV. Similarly, beta two is logit difference between TV yes and no. That is the TV effect for the kids that didn't get the classroom curriculum. Again organizing parameters in the way I've done here makes that pretty apparent. Beta two represents the TV effect when CC is equal to zero for those kids that didn't get classroom curriculum. Again the interaction represents how does a TV effect vary between no classroom curriculum and classroom curriculum, so beta three is the difference in the logit attributable to the interaction. We have the cluster random effect, so again, interpretation depends on how you code these variables, so it's very important that you code them and interpret them in a consistent way and the betas are just in there for the cluster effects. These are sometimes called cluster specific effects, these betas because they're adjusted for the cluster effect.

Here is a three-level model, so the previous model I wrote it as two-level. I didn't really specify is my cluster classroom or is it school. It could have been either one. Let's now write it as a three-level model. So here we've got- at level one we've got the schools. I'm sorry. At level one we've got the kids. Level one is the fastest moving unit because

that is the kids. They run faster than classrooms or schools. You can't catch them, so that is the kids. And notice here we got another subscript J or K. K goes from one to NIJ. IJ is important there. It's saying that a given school has different numbers of students. Then at level two we have the classroom, between classrooms within schools model, so the logit here for a given school comes down here to get modeled in terms of any classroom level variables potentially and now this coefficient, that is the classroom log odds comes down to be modeled in terms of variation between schools. So I put these variables here at the school level because again the different clusters, the different schools were randomized to either of these particular conditions.

So here is the three-level analysis without any individual level covariates, without any classroom level covariates. So beta not- Again, it's the logit for the control group. Beta one is the effect of classroom curriculum when TV is no. Beta two is the effect of TV when classroom is no. Beta three is the interaction and now we've two random effects, the classroom effect and school effect.

All right, a little bit about SAS for multilevel analysis of dichotomist outcomes. Good news. Good news. Good news, relatively speaking good news. There is good news. There is bad news. Which one would you want to hear first? I'll give you the bad news first because then the good news will see all that much better. This morning you used PROC Mixed to run for continuous outcomes. PROC Mixed has been around for a pretty long time and is pretty well- is a fine software program. As I mentioned it's much easier to estimate continuous outcomes than dichotomist outcomes, so we can't use PROC Mixed anymore, but SAS has two procedures to do it, GLM Mixed and NLMIXED, so that is kind of the bad news. You got to learn something else. You can't just use PROC Mixed. Now the good news is that between these two you can now use GLM Mixed. Why do I say that? Prior to version nine of SAS GLM Mixed did do this kind of modeling, but estimated it in a much more approximate manner and NLMIXED did what is called the full likelihood solution, which is generally what I favor and what many people in the literature favors. Now the problem with NLMIXED is not really it could be prefaced as an opportunity for learning, not a problem. It's kind of complicated to run actually, but it can estimate a lot of esoteric models, so we'll spend a little time on NLMIXED. It's much more difficult. We'll spend most our time on GLM Mixed, so that is the good news. GLM Mixed can be run for these purposes. It's pretty similar to Mixed. Not exactly, but pretty similar in terms of syntax. If allows for multiple levels of nesting, crossed random effects even. By default it does the thing that I don't like so much, what is called pseudo-likelihood estimation.

Again, computationally there is some challenges that you have to face in estimating these dichotomist mixed models. What pseudo-likelihood estimation does is it doesn't do a full likelihood solution. It does a pseudo-likelihood solution. It linearizes via sort of a **** expansion to avoid the integration over the random effects. What is this about? In producing the likelihood, so what is the likelihood? The likelihood is the function you're trying to maximize basically is exactly what in ordinary logistic regression when you produce those betas you're maximizing the likelihood. How do you get those? You

take the derivatives of the likelihood function. You set to zero. You solve. Typically you got to do the iterative solution to get to them.

Now in mixed models the models you've spoke of this morning the errors. The errors, the **** variable is normal. The random effects are normal. The full likelihood can be expressed as what is called the multivariate normal kind of format. You don't explicitly have to deal with integrating over the random effects, but in this class of models you do. That is the crux of how these different software programs are different for dichotomist outcomes. How they handle the integration over the random effects. Why do you got to integrate over them? You have to basically get a solution that goes over all of the random effects. It's a random variable, so in statistics how we get rid of a random variable often is integrating them out. If you have a continuous outcome that is normal, you got random effects that are normal that can be expressed with what is called closed form. You've got basically a multi-variant normal solution. You don't have to explicitly deal with that integration, but in dichotomist outcomes you don't have that because you're likelihood for your outcome, your dichotomist outcome it's not a normal. It's like a binomial and then you've got these random effects that are normal, so you have a challenge there how you're going to deal with that. What the pseudo-likelihood approaches do is they linearize it. They basically avoid this integration in a rather approximate way. Now people look at this. How bad are those estimates? And they found that they don't produce biased estimates if the number of level one or level two units is small and/or the ICC is large.

I'm going to try and put that into context. When I read that it reminds me I shouldn't be so hard on this today. Why? What is the difference between today and tomorrow? Today I got here at 2:00. Tomorrow I'll hopefully get here at 8am. That is one difference. Major differences today we're talking about clustered. Tomorrow we'll be talking about longitudinal. Now how does that relate with regard to this statement? Level one, level two, in our clustered scenario level one are like students let's say. Level two are like say classrooms or school. So what this method doesn't do so well if either these or these – You know these are rather small. Well so there is 28 schools. That is kind of smallish. It's not terribly small. There is 1,600 kids, lots of level one units, maybe not so much level two, so maybe that is a problem. ICC, is the ICC large? In clustered data, no, ICC is not large, so this issue of bias it raises its head more on longitudinal because in longitudinal you're not going to have the many repeated observations. You might have a lot of subjects, but you might- you're probably not going to have so many repeated observations and the ICC is certainly going to be big because the data are clustered within subjects and same thing. How correlated are the data within a subject? Within a subject the data are rather correlated. Within a classroom or school they're correlated, but not like within subjects.

So this issue of bias that the pseudo-likelihood procedures have it's really not so much of a problem today. Tomorrow yes and so I'll show some results that show the pseudo-likelihood, the default estimation in GLM Mixed and then this full likelihood estimation. I feel this way. If you can go full likelihood method that is better and these days with GLM Mixed you can do that, so that is what I would generally advocate is that if you can

do it why not? in the old days you couldn't do it so much, so then you had to go the NLMIXED and SAS because that was the only procedure that did the full likelihood solution, but these says you can actually do it in GLM Mixed and so that is a nice advantage. Now one thing about NLMIXED is that it only does two-level models. GLM Mixed does three levels, so that is kind of an advantage of that.

Okay, so we'll be starting with GLM Mixed today. Tomorrow we might have to get into a little bit of NLMIXED, but today we can certainly do all of our work with GLM Mixed. Now let's show an example and the new can actually try to estimate this. So it's going to read in a version of this dataset that we've been playing with today. Here is the version in which I created these variables and incidentally, I try to do this as much as I can as when I have the syntax is put things in uppercase that are SAS specific and things that are lower case or mixed are things that I have to name, so that you know what sort of like a SAS word and what is something that I have named. Now the variables that are in this dataset that you have available SID is school ID. SID, the other CID is not the you know. That's probably not a good choice calling one SID and the other one CID, right. One we could say in CID D. The other one is L SID I guess or something like that. So that is Class ID right. Actually I first has these as school ID and class ID and then it went past this and so let me rename those things, so that it fits on one line, so that is school ID and class ID.

Then I've got an ordinal version of the outcome T, tobacco health knowledge scale ORDL [ph], that's ordinal and then the binary version that we're going to be looking at right now. This is the binary version, so did the kid get three or more correct at post test? INT [ph] is a column of ones. For SAS you don't need that around. For some other programs you do, so it's in the dataset, but we can ignore this. It's not relevant for our purposes. Then there is a pretest, so how well did they do at pretest? We're going to potentially included that as a covariate and then the indicator variant of classroom curriculum, did they get it or not? It's coded zero if they didn't, one if they did, TV zero if they didn't, one if they did and then the product of those two.

So here is how to run PROC Logistic for ignoring the cluster. Have many of you used PROC Logistic? Do you know what the descending option is about? Let me tell you something very interesting. It's about statistical software history. Statistical software is an interesting topic. History is an interesting topic. You combine the two and bingo. You've got like something that is so incredibly interesting. No. Now remember when earlier I talked about the logistic regression model and then its convention to code Y, zero or one and typically you write them out on terms of the probability that Y is equal to one in the numerator. It's convention, but realize you could also write it the other way. You could have the probability of a zero response in the numerator and a one response in the denominator. It's a bit unconventional, but you certainly could write it that way.

Now prior to version five or so in SAS they did things the conventional way, putting- you're modeling the probability of a one response of the higher category. Then I think it was like five or something like that. It was awhile ago, believe me, but I remember when it happened. All the sudden I'm running a problem and I'm getting estimates that

are exactly opposite to what I'm expecting. I was like okay is my mind gone that bad that quickly? I went back to the dataset. I was all what is going on here? I go to the documentation. Oh SAS, thank you so much for now putting in the numerator the probability of the lower response category, so that is what they did. They made this change and the interesting thing about it is that when they made this change they didn't like make a big to do about this change. All of the sudden it was like it was there in the syntax or in the manual. Back then we had manuals, but it's not like they put it in bold or anything like that, so I often wonder how many results are out there in the literature that are exactly opposite because if you weren't careful you know. So I think they got some heat for that admittedly so and so they then allowed the possibility of changing the default, so by default if you run PROC Logistic and those of you that have run it might realize this, that you're putting in a numerator the probability of the lower category response, so if you code things zero or one it's the zero response. What happens then is that all of your betas are going to be of opposite **** then if you have a one response in the numerator.

So to flip that and sort of get the more conventional parameterization that is what the descending option is all about, so that is what that does. So I would say this too. If you're running logistic regression make sure you understand which way it's doing it because it is somewhat arbitrary though. There is convention there.

So here I'm running a simple logistic regression where I was stating my binary outcome and then my three predictor variables on the right-hand side. Now first I'll show your GLM Mixed, students in classrooms with a quasi-likelihood solution, so PROC GLM Mixed. No CL print, what that will do is I'm going to have to identify the classroom as a classification variable here. L SID here will be a classification variable. Now there is 135 classrooms and if I don't put no CL print on the output I'll get a printing of every classroom ID, so that is all that that does. The CL print will not then print out all of the 135 IDs for the classrooms.

And I'm going to identify that the clustering is attributable to classroom, so that is why classroom ID is identified here as a classification variable. Sorry, class there in SAS refers to classification. Then the model statement, notice it's pretty similar to logistic though not identical. The way in GLM Mixed you specific the descending option is to put it in parenthesis after your binary outcome and then after that you have to indicate the distribution of your outcome variable because GLM Mixed actually can run mixed models for a wide variety of outcome variables. It can do it for continuous. It can do it for ordinal, for accounts. It's actually more like a GLM sort of procedure, so here I'm saying distribution is binary because my outcome is binary and then this is the one that I always am amused about. Solution, by default you run this in SAS. It won't print out the beta estimates. You won't get the regression coefficients. You'll get tests for the different variables, but you won't actually the betas and therefore you won't- you can't do odds ratios because you can't exponentiate what you don't have, so I would have loved to have been a fly on the wall in that meeting in ****, North Carolina when they decided hey by default let's not print out the betas. That is what solution is. By solution you're going to get a printout of the betas. By default you don't get that.

Then finally the random intercept again. This looks pretty similar to what you saw with PROC Mixed and you can see that. Intercept here is a SAS keyword and then I'm saying subjects are clustered by CID, classroom ID and then type equals Cholesky. Let me explain that. What am I doing there? Yeah, it's not cholesterol. We don't like that. Take that off. No, it's the good type. It's not the bad type. It's the good type. It's the good. Which one is the good one? I'm a statistician here. I feel so embarrassed when I got to **** meetings and they're talking about all this stuff and I'm like I don't really know this stuff very well.

So this I know a little better. Type equals Cholesky. What that is, is this. You're trying to estimate here the cluster variance, so let's just go back to one of the equations since I like seeing those better anyway. So we're going to be estimating. Here we've got a class variance. We'll be estimating classroom variance. Now in a clustered scenario the variance is not going to be that big and in estimation realize this. A variance cannot be a negative quantity. It is what is called the bounded parameter in statistics because it can't be negative. It hits a brick wall, but you're estimating this by an iterative procedure. It's like climbing a mountain. You're maximizing the likelihood and so sometimes in the iterations it could go negative. Now that is going to produce bad results, so typically in statistics when we have a bounded parameter that we're trying to estimate it's a better idea to take some transformation of it that makes it less bounded. What Cholesky does is instead of estimating the variance it estimates the standard deviation. Now why is that? Let's say the standard deviation is .1. What is the variance going to be? 01, it's going to be a lot closer to zero, so the variance because it's the square of the standard deviation on a fractional value is going to be closer to zero where you don't want to get to than the standard deviation, so for clustered data it's generally numerically a bit easier, better to estimate the standard deviation rather than the variance per se. It's kind of it' not a very important conceptual difference, but it helps the computations.

Cholesky is a word- I think there is somebody out there named Cholesky actually. I think it was a Russian mathematician. Cholesky actually refers to the matrix square root. If you look in a book of matrix algebra and you see Cholesky factorization it's taking a symmetric matrix and getting the square root of that. Now here we've only got a scalar quantity, so it's just an ordinary square root, but that is why SAS has this word Cholesky. It really is it's more general than just- the way we're using it here, but what this will do is that estimate as a cluster **** parameter it will estimate as a standard deviation, not as a variance and again, it's beneficial to do that for computational reasons.

Now here is then the full likelihood version of GLM Mixed and there is only one difference here as far as I can see. On the PROC GLM Mixed statement, the very top statement I have one additional option, method equals quadrature and in parenthesis quadrature points equals 21. So this is strictly a 21 and over model. What is this? Again what I mentioned is with binary outcomes like this you have to explicitly deal with the integration over the random effects.

How many of you remember integration from your calculus days? Fondly I'm sure, right. What do you try to do when you're integrating? You're trying to get the area under the function, under the curve. What kind of curve are we integrating? It's a standard normal. We're assuming the random effects come from a normal distribution. So we've got a normal distribution that we're trying to integrate. What quadrature does is it does it in a numerical analysis kind of way. It replaces that integral with the summation over a finite number of points. So let's think of you were trying to evaluate the area under a curve. If you evaluate how high is the curve at say 10 points and use that you can get a pretty good fix about what is the area under the curve. You're sort of just breaking it apart. Instead of treating it as a continuous curve you're making evaluations at a finite number and then getting that as your fix on the area under the curve. That is what quadrature does. It replaces the integral with a summation.

Now think about this. You're trying to get the fix under a standard normal. The more points you use you think things thing are going to be better or worse? Better, better, right, better. Here I've chosen 21. To be honest that's rather a lot. Running these models for a long time typically if you go passed 10 it matters very little, but these days- and we used to tell people just use 10 or 11, something like that, but these days computers are so fast that we get even rather excessive. Here I'm going to 21. Now why? What is the problem with going more and more and more? Each piece of data that you have I'm going to have to make 21 evaluations of it. That is what I mean about the time factor. So for each of those 1,600 pieces of data at each iteration I'm going to have to figure out the area under the curve 21 times basically, so it can add a lot of time to the estimation. These days that is not such an important issue, so that is why I went up to 21.

Why odd you might say? Why not 20? The distribution we're getting over has a mean of zero. It's a standard normal. In the standard normal what is the most likely observation? What is the one with the greatest probability? It's centered around zero. If you choose an odd number of points you automatically have zero as one of the quadrature points. If you choose even it will be like five positive, five negative. It makes a very slight difference. I don't want to exaggerate with importance of it, but logically odd number kind of makes more sense because you get zero automatically in it and so I'm going excessive here with 21.

So that is how we can run a full likelihood GLM Mixed solution. Here is now another advantage you get out of GLM Mixed. With full likelihood you get a likelihood value, so if you want to do likelihood ratio testing you know compare this model to a model that has got three more variables in it you can do this under a full likelihood solution. You can't do that under a pseudo-likelihood and so that is another advantage of going the full likelihood route. Yes, you had a question.

[Comment]

Yeah, I would say 21 is certainly an ample, ample, ample, yeah. I'm sorry. The question was how many points to use. The more the better, but it will take a little bit more time. Here I'm going with 21. That is ample without a doubt. To be honest the number of points plays more into longitudinal than into clustered because again in longitudinal the variance, the person variance is larger. What that means is that the distribution is more spread out. That means to get it an area under that you need more points. For clustered it's not going to be that spread out, so you don't need so many points per se. You can even get- You can even try this yourself. See if you go with 5 points versus 21. It's not going to matter much for this kind of dataset.

[Comment]
The number of points? Again, what you're doing is to value- You have to do this integration over the random effect distribution and what you're saying is how many points you want to make the evaluation based on. So the more points you use the more precise of an answer you get, but the more time it can take to run. That is the tradeoff and again, typically in clustered sort of scenarios if you go beyond 10 things don't change much at all. I'm talking about changes like in the third decimal place or something like that typically. You're not going to see any great change.

[Comment]

Seconds, yeah, yeah, again, this- in the- I wrote software to do this kind of thing in the 90s. 21 points back then might have taken two minutes. I don't know, three minutes, maybe five minutes. Now it's not really that long. It is a little longer here with NLMIXED. Here shows the same code for doing NLMIXED, full likelihood solution. Now the moment you go from here to here to here I think you'll be glad that you can do this with GLM Mixed. NLMIXED is a really curious procedure because it includes programming aspects as well as your more normal procedures, so here if you look at this and I'm not going to go into this in great detail at this point, but I'm actually having to write out the probability statements, the log likelihood. I have to give starting values. I have to name all of the parameters of the model that I'm going to be estimating. So you might say why would I ever want to do this? There are some models that cannot be fit by standard software like GLM Mixed that can be fit with NLMIXED and which you- in the old days we would have had to write a Fortran program or do some complicated programming to actually estimate those models. For example, tomorrow I'll talk about for missing data, what are called sometimes shared parameter models in which you model the longitudinal process and the time until dropout. You model those concurrently. Those kinds of models have been in the statistical literature for like 20 years, but before NLMIXED people had to write their own code to do those things. There was no standard software to do that. NLMIXED admittedly it's more complicated than this, but it's a lot easier than writing out your own program with something like C or Fortran or whatever. So anyway, I don't want to get into it too much at this point where we don't need it, but when we need we'll see how to use it.

Then this shows a three likelihood solution with GLM Mixed, both the default quasi-likelihood on top and then the full likelihood on the bottom. Now here is something kind

of curious. I went to a three level and notice with the three level all you're doing is adding in two- you've got now two random statements and the way that you tell it with SAS that this is- this random statement is saying the data are clustered by school ID. Then this one is saying classroom ID nested within school IDs, so it uses a parenthesis for nesting. So this is school ID. That is at level three. At level two we have classroom ID nested within school ID, so that's how SAS indicates that. So I tried that with GLM Mixed with full likelihood and I went with 21 points. I thought it might take a little while to run. Well it didn't even run, so I went down, down, down, down. It will only do it with three level up to only one point. Now you might say that doesn't seem very precise. Well yes and no. Again, for clustered it's not going to matter too much and in fact the quadrature solution with one point is what is sometimes called the laplos [ph] approximation as well. For clustered it works reasonably well. It's fast. For longitudinal is where I wouldn't be really satisfied with this approach, but for clustered, for our purposes today that is actually not bad. It's nothing to worry about.

[Comment]

Yes, yes, right. Good question. So where do you get the starting values in NLMIXED? Because here what I'm putting is I'm putting a logit estimate for the intercept, a logit estimate for the classroom curriculum effect, for the TV effect, for the interaction effect and then for the random effect standard deviation. That is what I'm putting on this statement. Where did I get those values? For the betas what I did was I took the betas from the ordinary logistic regression analysis, from PROC Logistic. In NLMIXED you have to specify starting values for every parameter. Where do you get them from? You estimate a simpler model. Use those estimates. It's going to put you kind of in the ballpark. It won't be exactly right, but it will be somewhat close. Now for the random effect standard deviation I put one. Why did I put one? It was easier to write one. No, here is the logic. I thought of in terms of an ICC. Remember in this model the ICC is the cluster variants divided by the sum of the cluster variants and Pi squared divided by three, so let's think of Pi squared divided by three. That's about three. It's a little bit more, but it's about three. So with a standard deviation of one what I'm saying is that the ICC is one divided by one plus three or a quarter. For me that works pretty well for both clustered and for longitudinal, for longitudinal typically the ICC is going to be greater, more like .5, .6. For clustered it is usually going to be a little bit less, like 5%, .05, .1, so I often just set this thing up with that value because it's kind of in the middle and it's kind of close to both, but that is basically where I- What I would say is if you use NLMIXED don't ever put as a starting value for a variance parameter zero. That is not good because the estimate has to be non zero, so you want to certainly put a positive value there, but that is where I put that.

[Comment]

Not necessarily. So if the question is if you put something very bad, something way off would you converge? No, not necessary. That is the thing that- the other challenge with running NLMIXED is you want to put- Now I don't want to scare you too much. I run this MLMIXED on a lot of these problems. A lot of times I'm lazy and I put zeros for

the regression coefficients. I can’t say that I have ever run into a problem there. Where I have run into some problems with convergence is when I'm estimating some more esoteric models than this. Like for example, when I'm- maybe little shared parameter models where I'm concurrently modeling longitudinal and time to dropout. Where you’re doing something a little bit more advanced than something like this the starting values can play a greater role than this, but yeah, there is no guarantee that you’re going to go to the right solution per se.

[Comment]

Sure, time for a break, but I just was getting started. Okay.
[Comment]

Okay, so we’ll convene again in what? In 15 minutes. 15 minutes, okay.