Multilevel models for clustered continuous outcomes, part 2—2010 MCH EPI pre-conference training workshop

SUBHASH ARYAL: The file name is- Let me now figure out what is the file name. The file name is going to be TVSFPmix, so if you go the open window over here on the sign into new SAS data exam **** you will see this TVSFPmix. Let's open it. Now the first thing that we need to do is- Now let me ask this. How many of you are very comfortable with SAS? Okay, almost everybody is, so I'm going to go over some of these things like what is the data and file import, so some of it may be repetitive for people who are already very comfortable with SAS, so bear with me and then we'll proceed to the regular analysis. So in this case this is the name of the dataset, SAS dataset that we are creating and the in file statement is basically to tell SAS where our data file is and in this case it says ****, so now we have to change this to identify the part where the place is, so for that one let me just put it over here because everybody has their data in training, so let me copy this and I'll show you what is the exact part. Most **** are going to be the same thing for everybody.

Now if we change it to C training, San Antonio **** then you should have your correct part. Instead of C mix just this thing. The data will already be there.

[Comment]

Okay, so in this case- Okay, this one I just changed it, so don't worry about this portion then please. Just put C training and then this TVSFP2b. So don't worry about this middle portion that I've highlighted, just C training, so now let's just run this thing and see if we are going to be getting anything, so let me just run this portion and see if the data is there. So let me put a run statement over here.

And then the next one is the input statement. What the input statement does is basically it says that what are the variables that is in our dataset, so in our dataset these are the names of the variables that we have and that is what ****. The first one is the school ID. The second one is this class ID and then the post score. Basically there is a separate intercept that has been created, which is the ****. That is a column basically of one. Pre TSKS is the pre score. CC is whether they got the- what kind of curriculum they got. TV is whether they got TV or not. CCTV is whether they got both the classroom curriculum and the TV, so let me just run this portion to see if my dataset is- if SAS can find my dataset.

So let's look at- and if you run your ****, so if you just add a run statement like the way I've done and if you just run it you should be getting all of your data in there. Is everybody with me now? Okay, now this **** statement it's basically to beautify the output. Instead of this being zero and one if we want CCTV, those things to appear in your output you want to put the format

statement. Those of you who are not very comfortable or familiar with SAS my solution would be don't worry about the format statement because you may get error and it won't run, but it's not going to do anything for our analysis, so it's there. Let it just be there, but basically what this is doing is when it says **** it will print out in the output that no CC one is yes. If you don't put a format statement it is just going to put zero and one in your output.

So now let's just click the running man and then run all our analysis. And it may take some time because this mixed affect model sometimes take a long time to run, so let's just hit that running thing and then get out our analysis. Did it work for everybody? Did all of-? Okay, is anybody having any problem with this thing?

[Comment]

Warning on the very- Warning, the variable type **** in the data says that is not ****. For right now let's not worry about this. When it comes time we'll worry about it if we need to worry about it. Sometimes SAS will do something and then- there is a- Okay, last line of code. Let's see what it is. ****, aggregate classroom analysis, CCC. Okay, we'll come to that when the time comes, so.

So in this case the regular regression analysis. In this case if you look at the SAS program it says "student level **** analysis ignoring clustering", so **** is the simple linear regression analysis that we are doing and then we don't have anything. All of us who have run SAS in this model this is basically a simple linear regression model where the outcome variable is a function of this variable, the pre score and the different level, different kinds of intervention data that student got, so let's look at that.

Now in this case **** is the method that we use in SAS for regression analysis and whenever we are doing that we need to specify what is the model and our model was that the post score is a function of these things. We had the pre-score and then the values of the CCTV or whichever kind of intervention that they got. So let's look at-

[Comment]

No, when we put CCTV that is the- They got both of them, right?

[Comment]

Group, they got- Okay, let me go back to the slide.

[Comment]

Yeah, in this case whenever you see this CC times TV interacts then this is basically the group of people who got both the intervention in that one. So in this case the number of observations used is 1,600, so if you look at some of these parameter estimates you will see that one of things that you want- we want to look at is R square, which is it's basically 11.70 ****. It's explaining. Our model is explaining almost 12% of the variability in the data and if we look at all of these

parameters- the significance of these parameters we see that at the .05 alpha level all of them are significant. All of them are significant. Let's go to this thing, the notes that we had and let me see if I can- See, if you look at over here the intercept in this case whenever we are doing the student we are ignoring the class level and we are just doing at the student level, so in this case it's 1.66613, 1.66126, which is basically significant and then if we go here. Let's see now. So in this case whatever is the score for the pre score it's 32518 is 3252, so it has just been rounded and we see that all of them are significant. In this case whenever we are doing the regression analysis all of them are significant for the analysis. This is just doing the student level analysis ignoring the other and whatever it is in the bracket it's the standard error, so to get these T values you are basically dividing this parameter estimate by the standard error and this is what SAS is producing, so at the .05 level all of the variables that we are starting, including the intercept is significant.

Now let's move from this one to the next one, which is the one that we'll be using for the mixed affect analysis. Now in this case we are again using instead of **** regression we are using **** mixed, student level **** analysis ignoring clustering. We are doing the same analysis that we are doing over here, but we are getting a different kind of estimate. This is the **** estimate, the least square estimate and this is the maximum likelihood estimate. It's two different techniques obtaining the estimate. Now whenever we are doing- getting this **** this is what we call an unbiased estimate. The maximum likelihood estimate for our regression analysis, these are biased estimates and this is what we are getting from this one.

Our main goal is not even this analysis. Our main goal is to look at this portion of the code, which is basically our mixed affect analysis. Now for this one, to do this mixed affect analysis we'll be using the **** mixed. Good thing about SAS is a lot of these things that the **** is similar to whatever the kind of model that we are using, so for regression it's **** and for mixed affect models it's **** mixed, so we'll use **** mixed ****. Now the default method in SAS whenever you are doing the mixed affect analysis is called the restricted maximum likelihood method called REML. Now if we do not put this statement that says meta **** to ML then you are going to be getting that REML estimate. You can use REML estimates in some situations, but usually whenever our goal is to look at the fixed affects if we are- just like the way we did it in the previous slide where we're comparing whether this variable is significant, that variable is significant for the fixed affect in those cases using the REML estimates is not appropriate because that adds us towards the number of fixed affect, so while doing this analysis you have to remember to do the- put this statement that says **** to ML, so that you are obtaining the maximum likelihood estimates and not the restricted maximum likelihood estimates.

And the final thing here that we have is the **** test, which is the test for the core variance and basically what this SAS does with this **** test option is it does the **** statistics where **** statistics is the estimate of the variance divided by the standard error, so that is a **** test. Now while doing this analysis it's not recommended that you use whatever is the output from this one. This is just for like your preliminary thing to see whether you want to do it or not. The proper way of testing these variance components is the likelihood ratio test that we just did before. This one is going to produce what we call the **** test and **** test it's not recommended that you do the **** test. We will do the likelihood ratio test for the analysis.

Now this is the class statement and basically what the class statement in SAS does is it classifies the variable. It gives different classifications, so put if you have like let's say race and if we divide the subjects into let's say African-American, Caucasian and other, so we have three different groups. Now if we have three different groups to run an analysis in statistics you have to do what is called the dummy variable. You have to create codes for those things. Now what you can do is like you can create the code itself of if something is a classification variable that the variable just denotes different class you can put that in the class statement and basically what SAS does is it creates the classification variable for you. Now if you're just 2, 3 levels you may be able to do it on your own, but suppose there is a variable with 10 different categories means you have to create 9 dummy variables. Now that is going to be a very painful process, so in that case we'll just use the class statement, so basically by putting this class statement we are saying that the subjects are classified by ID. Now the model, the way we write the model for the fixed affect is exactly similar to the way we do it for the simple linear regression model. We say model where our model is what? The post score as a function of the pre score and then these are the other three core variants that ****. These are the intervention things whether somebody got CC intervention, TV intervention or CCTV. Yes.

[Comment]

I'm going to come to that. I'm going to come to that. Now basically what SAS does is if you don't put this solution statement it is going to run the analysis, but it is not going to print out the parameter estimates and the standard error and the T test or the Z test, whatever it uses while doing the **** mixed, so by putting slash is equal to solution you are basically asking- telling SAS I want to see the output from this analysis also, so after the model you may want to put this slash solution statement so that you will see the output from this analysis.

Now in this case we are going to be running a random intercept model with a random class thing, so for that one what we are saying is we want to put a random statement for that one and if we just want to run a random intercept model the keyword is intercept, so random intercept. Basically if you use the statement random intercept SAS is going to understand that you want to run a model, nested model that just has a random intercept. Now where are you going to be putting the random intercept? For all the classes, you want to give a separate random intercept for all the different classes, so in that case you want to put subject is equal to class ID.

Now if you are doing a random intercept for the subject then you would put class is equal to subject ID and sub is equal to subject ID over here, so wherever you are putting the random intercept this is where you are going to specify. Why don't you take a look at it for a few moments and then if you have questions let me know.

Now let's run this portion. If you have already run this analysis if you go to- Okay, let me do it again and then what we are do, we are going to delete all of this. If you go to the- Okay, now just a moment. Give me a moment. So basically if you go to the results and then right click in the results and then it will ask you if you want to delete it. You can delete whatever is already there. Let's delete it so that we don't get confused with all the output. We are just going to keep whatever we are running, so if you click on the output you may get an error saying that this **** is something there. Let's see. Let me do this again. So this is my output. I go to results. I right

click. Let me just do delete. Yes. Now if this happens just click okay and in this case let's highlight the SAS code that we have here and in this case if you go to this button over here that says "break", click on "break", then it will give you this popup window and you say **** data ****. Okay, now if you right click on this REZ and then delete it SAS will allow you to delete it. Basically this is an interactive process, so unless you tell it to quit it won't quit.

So now since our interest is just this portion let's just put a run statement over here after **** mixed, the score that we just did. Let's just add a run statement over here. Let's us put it like this, the run and then let's just highlight this portion that we are just done and then hit the running person here. Is everybody with me until now?

[Comment]

Yeah, if you put the run after the title statement put a semicolon please and then let's just highlight just this portion so that we don't get confused with all the output that we have over here. Let's just highlight this and then hit the run statement here.

[Comment]

Okay, if you want to delete the results if you right click on it you'll see the delete statement. Click on the delete statement and then it deleted everything and just that **** is there, right, so in that case if you go to the program that we have and if there is a break thing over here on top, if you see after the cross. Okay, let me see. No, if you highlight it here. Just click over here somewhere. Now it's there. Yeah, no problem. No, you have to click on this button right here, right there. You click over here and then if you'll see that **** second button, if you hit okay, okay, two times and then if you go to the **** and right click and then hit delete it will delete it now. Everything is gone, right.

Now let's go to the code that we have that we are going to run, which is basically this one, two levels, student nested within classroom analysis and then go to the- let's put a run statement after the title just like the way I've done. Let's just put the –r-u-n over there and the semicolon. Is everybody with me until now?

[Comment]

It's not finding this TVSFP data.

[Comment]

Okay, then let's- Did everybody delete everything that we wanted to delete? Okay, let's run everything from here then, this portion. Let's highlight this data and this **** format statement. Yes, highlight this **** format and the data thing and then let's just run it. Now our dataset should be there. Are you getting that no data found error? Once this is done now let's go to the **** mixed now. I may get a no data found error, so. Let's highlight this portion, **** mixed **** ML **** test and then let's run this part. Now did everybody get this where it says mixed analysis now? No, let's not get frustrated. We will get it there. We'll get it there. It happens.

So many times I've felt like throwing the computer out of the window like oh, what SAS is doing, so we'll get through it. So okay, to get through that **** let's look at the top portion now.

[Comment]

Highlight just this portion. Now we're using SAS. When I was doing my thesis I was supposed to run some nonlinear model. It took like maybe four days. It ran, ran. I just left it, went home, come back tomorrow, look at it and after four days it just crash saying that there is no result and I have to run the whole thing again, so sometimes it's a very frustrating experience, but. So did everybody highlight this and get the dataset?

Now let's go to this **** mixed **** to ML. Let's just highlight this portion and run and usually if you just run this portion you should be getting your analysis. Now let's look at these things. Now basically what SAS does is like it says dataset. Now you want to verify that you are using- doing your analysis. See the problem with SAS is here we have not told SAS which dataset to do, so basically it will create a dataset and whatever is the last dataset it created it runs the analysis on that. Either you have to explicitly tell SAS that I want to run my data analysis on this dataset, if you don't do that it may be a good idea to see where it is. Now if you see- Did everybody get this output? Okay, dataset is this thing. The dependent variable is post ****. Now this **** instructor, it's some **** instructor over here. Subject affect is class ID. We are going to be using the class ID. It's the class that are defining whatever is the nesting thing. **** estimation is ML because this is what we want to use. We can the REML estimation if our interest is in the variance components, but most of the thing that we are going to be doing is just for the fixed affects, so whenever we are doing the analysis for the fixed affects we are to make sure that we use the ML and how do we get this ML is we have to specify in SAS meta **** to ML. If you don't specify it in SAS it is going to do something called the REML estimation and we don't want to use REML estimation for this kind of fixed affect analysis that we are running. And then we don't have to worry about some of these things, whatever SAS is doing on the background.

Now in this case if we have this **** we have 135 classes and these are the ID for all of those classes. And then what is the data dimension? Let's not worry about these things. It says that the maximum observation possible is 28 and we know that all of those things, so let's not worry about some of these things. Let's look at the analysis thing. Now in this case it says number of observations read is 1,600. Observations used are 1,600. Sometimes it will read some observations and the number of observations used will be less. If there is missing data for any of those variables suppose you have this one subject going like this and you don't have the information for CC means it cannot use that information. The thing about mixed different model is it can handle missing data. When we say it can handle missing data is suppose we are measuring somebody for like six time points and we just have information for the five time points it can use the data, but if we don't have- within one line if we don't have complete information it obviously cannot do it and that is what sometimes happens and now in this case you'll see something called the iteration history. The reason that iteration history is here is that ML **** the maximum likelihood **** that SAS uses over here is an iterated process. It's not like a simple linear **** estimation where you get what we call the closed form solution, so this

one we don't want to have to worry about it. What we are interested in is definitely this statement where it says convergence criteria met because if you don't see convergence criteria met sometimes it will put an error estimate that says did not converge ****. The **** whenever it was going through this it's an **** process. Whenever it is going to this iterative process it just could not find the solution, so we always want to see this convergence criteria made.

Now in this case the next thing that we have over here is the core variance parameter estimate. What this core variance parameter estimate does is it gives us the variance for the things that we have, so in our model we have what? A random intercept, so it's going to give us the what is the variance for the random intercept and then what it its standard error and what is the residual variants. This is an estimate of the *****. This is an estimate of the variance. It's not whatever is the error. This 1.60 is not that the error is 1.6. It's the variance of that error is 1.60 and in this case the variance of the random intercept is 0.08. Now in this case if you use- This is what we call the **** test. By just putting this **** test option you are going to get this Z and the probability of Z. This is called the **** test and **** here test if you look at these numbers it's telling you that the random intercept variance is significant, the residual variance is significant, but my suggestion would be what is- put in Dr. Hedeker'sbook and some other books that I have seen is we can- it's better to use this only as a guiding information rather than as a test. To see if any of these things are significant or not we may have to use the- the better option is to use the likelihood ratio test, so yeah, from this one we see that they are significant, but let's not use this. We are interested in this estimate, but not in this test. These are just for us to guide us and we'll be testing this formally using a likelihood ratio test.

Now in this case SAS will produce instead of the log likelihood that we saw in the previous slides that we had, SAS will itself produce the minus two log likelihood that we want to use for our likelihood ratio test and if the models are not nested in those cases we may want to us the EIC, EICC and the BIC and SAS gives the notice on how to interoperate it. Whichever is smaller than that is the better model. So we will be using these kind of- this information over here.

Now our interest may be more in this portion because these are the things for the fixed affects. Now when you are looking at the fixed affects previously what happened and the reason that we put that slash solution statement was if we did not put this statement this thing would not be produced. If you want to look at it we may want to put the slash solution thing, so once you put that slash solution we'll see that. If we look at the output we'll see that only the CC is giving a classroom curriculum as a significant affect. Existing for that one the TV and the CCTV, which is basically both the curriculum and the TV. That intervention is not significant. Did everybody get this screen, all of you?

In this case now previously if we looked at the ordinary least square expressions and the maximum likelihood we saw that all five of them were significant, but once we took into account the clustering of the data and ran a clustered data analysis we see that only this thing is significant. The other ones are not significant. Let's look at our slide and see. 1.6776 student in class, so if you see over here in the student level analysis that we did before this one intercept, pretest, classroom, television then whatever we call the interaction basically it's the interaction. When we say interaction it's the both of them together. That one all of them are significant, but

if you look at the multilevel model you see that only these three things are significant. The other two things are not significant. So by doing the analysis in the multilevel approach we see that this is the correct- this is the significance is here, but for the other two it's not significant and then if we look at this class level variance. Adding this class level variance does it make a significant improvement in the model? Yeah, it makes a significant improvement in the model.

Now how do we say whether it is significant or not? We have the log likelihood for this model, minus two log likelihood for this model. We have the minus two log likelihood for this model and if we do a likelihood ratio test like the way I saw it before where you are subtracting the minus two log likelihood between the two models you will see that this is going to be significant and this is the estimate and this is the standard error. So by doing the same analysis in two different ways where here we basically ignored the clustering thing, here we have said that the cluster is at the classes level, so by doing this analysis like this we saw that the significance has changed from everything being significant to only this one being significant, the other two being not significant.

Now let's run the next model. Now again let's go to mixed, right click, do delete. Are you sure you want to delete? Yes. Now let's- Once you have deleted everything let's go to TVS, again the same program. Now let's highlight something else. Before we ran this analysis where we said that subjects were nested within the classroom, now we are going to run the same analysis by saying that the subjects are nested within the school, so- Yes.

[Comment]

When you are- The variation in the cluster means by putting this cluster affect basically what we are saying is everything within the cluster they are similar. Those are correlated and then given those things then the clusters are independent, so by putting this random affects what we are doing is we are saying that all the people in the same classroom they share some common properties and we are taking care of that common properties that all of those people are saying. If it's the same school, same class, whatever it is randomizes and basically makes your groups comparable. When you randomize it's the main purpose of randomizing is to make sure that you are giving them the- you are not introducing any bias. The groups are similar, but in this case even the similar groups, whatever you are saying for your treatment randomizing is not for the cluster. Randomizing is for this CCTV. You may say that whenever you are doing your study if you randomize- suppose you like somebody you may say that I am going to give you classroom curriculum is the best. I am going to give you classroom curriculum. This classroom and TV is the worst. I'm going to not give you- I'm going to give you this, so to take care of that bias you are going to do a randomized study. You are going to randomize people into different treatment and but in this case what we are doing with this multilevel setup is we have randomized all the groups are like that, but within each cluster whenever we are randomizing we randomize everything to the same school. The design is that everybody within the same school got the same intervention and in this case when we are analyzing the data we have to make sure that we are saying that the students in this school are same. They are related to each other. If you go within even a same school district there are some school who are considered really good and there are some schools which are considered not so good, so in this case there may be a big difference in the socioeconomic status of the people who go to this school compared to this school. Maybe

the reading level, everything may be different, so to take care of the fact that all of those people in the same school are similar to each other we are putting that random.

Now in this case what were are going to be doing is like instead of the class level thing we are going to do our clustering at the school level, so everybody within the school, so in this case let's again put a run statement over here just like this and again, see the difference between the previous case is the classification variable was the class ID. Now the classification variable is the school ID, so we are saying that all the people within the school are nested. So if I highlight this person now I may get an error, so let me see. Why don't you try it also and see if you get any error or not? If it comes we'll have to do- So once we do this analysis now again the same thing if you look at the model information. Now see, we did our analysis now at the school level. There were 28 schools and that is why the school level has only 28 clusters now. There is 28 clusters. These were the IDs for the 28 clusters. Before there were 135 classrooms, so we were ignoring the schools and we were taking care of the 135 classrooms, so that is why we have a different number now. It's at school level, but still our observation is 1,600, 1,600.

Now again in this case you'll see that this an estimate of the random affect variance. This is the standard error. This is the residual. Again, if you look at this value you'll see that the random intercept at the school level is statistically significant, but again, let's only use this as a guiding thing, not as the final result. To see if the random intercept is significant or not it's recommended that you do a likelihood ratio test and before looking at this, this is the key thing that you want to look, convergence criteria met. If the convergence criteria does not meet we are not going to be getting any estimate for this ****.

So now go ahead and let's look at the fixed affect. If you look at the fixed affect now we'll see that those three things that were identified previously they are significant. The TV thing is not significant, but the CCTV people look at both the classroom curriculum and the TV. It becomes marginally significant. It's slightly above .05, so it's marginally significant over here. So in the ordinary regression analysis everything was significant. When we did the analysis using the class level variable only these top three variables were significant and now when you are doing the same analysis using the school level analysis this thing became significant over here.

Now the question again comes like when do you decide, how do you decide whether you want to do it at the school level or whether you want to do it at the class level, so in that case that is part of the thing as you as a researcher that you will decide how you want to do it. If the classrooms within the schools are very similar and you just want to put it that random intercept to account for the intercept between the schools then you may want to do it at the school level. If not, you may- if you want to compare the different class levels in that case you may want to do your analysis using the previous thing that we did, which was the class ID thing and again, only the difference is if you go and look at over here you'll see that students in school that this is significant, this is significant, this is significant and then the interaction whenever, what we call is the interaction basically that is the thing that is where people got both the CC and the TV. That is significant at the 10% level. **** is **** less than 1.0 and if from the SAS output we saw that. The P value for this one was .0669, so by doing our analysis in different ways we are coming up with results that are different from each other and if you look at the parameter estimates those parameter estimates are also different. Here all of them are actually negative, but

compared to here it has decreased and here it has increased and if you look at this parameter again, 64, 63, 66, so the parameter estimates are differing and then the things are differing and in for the school level variance if you want to do the school level variance if you do a likelihood ratio test between the model that did not have the school level variance and where you have the school level variance if you do the analysis you'll see that that one is also significant.

Now for all of these things while doing the test I can use these values for the P value and say whether they are significant or not, but when I want to do something for this variance component then I should rely on the likelihood ratio test rather than the **** test that SAS produces by default, so we are doing this. Now suppose instead of just doing this we said that there is a random classroom affect that was significant, random school affect that is significant. Now since we have information for the classes, we have information for the school let's take into account all three levels of nesting, that the students are nested within the classes and then the classes are nested within the school. So we have gone from a- Yes.

[Comment]

No, we are looking at to say that the- the question is that how do we know that this class intercepts? They intercept for the class of the school is significant. For that one we are not looking at the fixed affect. We are looking at the class variance. There is a variable called class variance, which is the variance for the random intercept. See if you go back to this model over here, here we put a random intercept at the class level. Is that random intercept significant? Here we put a random intercept at the school level. Is this random intercept at the school level significant? How do you decide that? For that one we are going to be doing a likelihood ratio test between the model using this minus two log likelihood from the regression model and this model and once you compare the likelihood value with **** square with one degree of freedom it's just one extra parameter. When we do that likelihood ratio test we will find that these variables are significant and that is why we say that the class level intercept, random intercept at the class level is significant because this is significant and the random intercept at the school level is significant because this is significant.

[Comment]

No, it's not. SAS won't produce these things by default. You have to do the likelihood ratio test. Just to get an idea about whether it is going to be significant or not you can always use this core variance parameter estimate where it say intercept. This is significant here for the class level thing, but the proper way is not looking at this core variance test. We are to look at the likelihood ratio test.

Now we found the class level things significant. We found this- Yes.

[Comment]

For this thing- The question is to do the likelihood ratio test which models are we comparing? So for this one if you want to look at the likelihood ratio test for the random intercept at the class level then you would be looking at this thing for mixed affect model that does not have the

random intercept. With the model that has the random intercept at the school level, class level. if want to compare that thing for the random intercept, regression model with the random intercept at the school level you are going to compare the minus two log likelihood value for this model compared to this model. So basically how is the likelihood ratio set up? You have a model without the parameter. You have a model with the parameter, so you get two different values of log likelihood value and then you do a subtraction of the likelihood value to see which one is significant.

Let me give you a few more minutes to digest all of these things and then think about it and if you have questions you can ask. Let me get some water and then we'll again start. Yes.

[Comment]

Yes, exactly. Yes, television and those people-

[Comment]

Yes, yes.

[Comment]
So basically this thing with the classroom intervention is like when you tell the students in the class that do this it is making an effect, but the same effect, if you are telling them something and if they watch TV also it's going away, so in this case when you take into account the nesting that means all the people who are within the classroom, so those people who got just this classroom and this was a randomized study that was done at the school level, so in a particular school all the people got this, only this classroom curriculum. In another school all the students only got the television, so in that case basically the significance, those things, whatever was the post score for people who only got the classroom intervention; those things were different from the students in other schools where they got only the TV, both or neither one of them. Yes.

[Comment]

Okay, the question is how do you decide whether this is class and this is school? So in our case what we have done is we are right now just doing two-level model. In this case we ignore the school level thing and we ignore the class level. In the first case we ignored the school level thing. We just did it at the class level. And in the second level we ignored the school level thing, the class level thing and just did the school, so since in our data it's a three level design, so in this case basically the level of nesting is what? The people, students, class, school, the highest level of nesting. The first analysis we ignored that one. We just did this little lower portion. In the second one we ignored this portion, just did the first ****. Now in the three-level model we are going to be comparing all three of them, so when we are comparing all three of them if we find that this one is not significant at the school level that means it may be okay to just to the class level or the student level depending upon what questions you are trying to answer, but in this case since there is a possibility for us to- we have the data for all three levels of nesting, the biggest model is going to be doing the highest level of nesting and then come out. That is basically going to what usually you may have already done in regression analysis thing. You can

start with the backward elimination process, which is the saturated model. You may put everything in your- all the variables in your model and then check the significance of one at a time and whichever is not significant you kick that out. You keep the rest of the thing. That would basically be doing the three level design, see if the three level, highest level of nesting is significant. If it's not you come down to the lower level of nesting.

Now the other approach is the forward selection kind of thing where you are looking at one variable at a time and then proceeding forward and in this case if you just do the two level and then proceed to the three level then you will do the two level analysis and then you will go to the three level analysis to see if the three level analysis is significant. Now while doing this analysis if I'm doing this analysis I'll be just doing the three level data **** because I have all the information. Why not utilize all the information for that one? Now if I see that the school level variance nesting at the school level is not significant then I may come to a simpler model because simpler models are always easy to integrate, easy to understand, so in that case I may be doing my analysis in that way.

Now let's go to the three-level model and see if taking care of all the nesting structure- You have a question over there.

[Comment]

You mean in SAS.

[Comment]

You mean this one. In this case again, the regression analysis whenever you are doing the ordinary regression analysis also, when you just do the class level **** you are basically saying that there are how many data points, 120 data- all the subjects within that class. In this case whenever you are doing it in the student level analysis means you are talking about saying that all of these are the student level. For right now don't worry about the class level analysis. Let's just look at the student level analysis where we are saying that all the students are independent. In the class level analysis it will be all the classes will be independent kind of thing, but for right now let's look at the students. All the students are independent. The students within the classroom are related. The students within the schools are related and then the students are nested within class. The class are nested within school and that is what we are going to be doing with our final **** analysis. Yes.

[Comment]

Okay you mean the likelihood ratio test. Okay, the question is which models are being compared to the likelihood ratio test? Let's look at this first one. So the value is 533- 5377.9, so this is for a model where we are running just the student level ML analysis, ignoring the clustering, so this is just the clustering where we are saying this is just a simple linear regression thing. We are estimating the maximum likelihood estimate. To do the likelihood ratio test you have to get the ML estimate, so can somebody please note down this number? What was that output? The value was 5377.9, so let me go to my calculator over here. 5377.9, now let's again run the same

thing for this class ID. 5360.0 minus 5360.0, so the value is 17.9, so 17.9, so how many extra parameter that I'm estimating between these two models? One extra parameter, random intercept. Now if you are not comfortable about thinking how my parameters are there the easiest way to look at it is let's go the information over here. You will see that there is one parameter over here for the residual in the previous model and there are one, two, three, four, five total, six parameters in the simple linear regression thing and if you go to the mixed affect analysis and if you look at this after the convergence status you have one parameter, two parameter, five and two, seven parameters, so the difference in the number or parameters between the models is one, so now it follows **** square distribution with one degree of freedom and in this case this is 17.9 **** square with one degree of freedom at 95% level or **** .05. There is 3. ****, so that is why it's significant and that is why we say that this model is significant improvement over the other model. Yes.

[Comment]

Yes, you will be- In a multilevel model you may be able to look at something further **** individual level variable. Some of them that will only be at the class level and you may be able to look at the interaction between those kind of things. When you put all of those things then basically this class level variance may change and that is going to affect the standard error for some of these things over here and when that happens you may get a significant affect and that is the importance of choosing correct variable, what you are studying. In a properly designed study we always think about what variables we need to be choosing and then putting them at the correct form so that everything that is **** going to these variances over here that is explained so that it reduces the variance and we may get a significant result. Yes.

[Comment]

No, we are testing to see whether this is going to be important compared to this one. Now to see this is a three level design, do we need a clustering at the school level and the student level? We are going to be doing this model, three-level model and compare it to these models and see whether that is going to give us a significant improvement or not and that is what we have ultimately done here in the final thing I think.

Now again, the same thing, **** mixed **** to ML **** test. Now some of you already pointed that out. If SAS gives you an error saying that it did not find dataset the easiest option would be to tell SAS explicitly where you are running the analysis, so in this case if you just add data is equal to whatever is the name of your dataset over here because this is the name of my dataset and I usually, whenever I'm running it I try to put this because sometimes in the intermediate steps in SAS I may create some dataset and if I don't put that data is equal to this thing in SAS, SAS is going to analyze the different data than I thought that I wanted to analyze, so in that case it may be a good idea to put it like this.

Now we are running a three level student nested within classroom nested within school analysis, so in this care our classification variable, our **** class level variable and the school level variable. School **** name of the school. Let's say this is school number one. Within in school number one, class number one. School number one, class number two, so we are going to

be getting our data like this, so in this case model is again, the fixed affect part is the same thing. We have already decided what fixed affects that we want and in general whenever we are talking about public health study in general these fixed affects are the things that we are interested. It's not the random affect that we are interested. We always want to make conclusion about the fixed affect, but while drawing those conclusions on the fixed affects we have to take into account the random affect and that is what we are doing for our analysis. Now there may be instances where you are actually interested in the individual rather than the group. For that one we'll opt in something called the empirical **** estimate that tells about what is the intercept for a particular person or a particular class, how much is each class different from the other things. We may be able to do that also, but in most of the cases these are the fixed affects that are important.

Now the first level of nesting is at the highest level of nesting, which is the random intercept for the school thing, so we do a random intercept for the school level analysis and the second level of nesting is we want to put a random intercept for the class and when we want to put the random intercept for the class we are saying that the classes are nested within the school. By putting this thing in these brackets like this we are basically saying that- we are telling SAS that we want to do an analysis where there is a random intercept for these classes, but these classes are not independent of itself, that each of these classes are nested within the school, so this takes into account the three level nature of our design. So let's run this thing and see what happens.

Now if we look at the results for just- Okay, let's go start actually from the top. Now see, if you look at it because we have two different levels of nesting SAS is saying that for the student level, class level there are 135 of these levels, 135 classes. There were 28 schools. This is the information for the 28 schools and now rather than looking at all of these things let's figure out whether this thing is convergence criteria met because we always want to look at this thing that says convergence criteria met. Now when we are doing this whenever we start some of these complex models sometimes what happens in SAS is SAS will force something. Basically what we are trying to do is like we are trying to estimate all the information based upon our data. We want to get an estimate for all of them. Sometimes SAS cannot estimate. The method cannot estimate all of those independent- information independently, so what SAS will do if we were to like set something at zero, set something as one and then estimate the rest of the parameters, now that approach that is not going to be correct for our analysis. To figure that out what we want to do is like we want to look at the log file also, so if you look at the log file if you just see convergence criteria met and then you don't see any warning message we are fine, but sometimes you'll see some warning message where it says that it will say that the geometrics is not positive definite or it will say that the final **** is not positive definite. Basically there will be a warning message over here. If there are those kind of warning messages even if your model is converging it may not be- you should not be using the parameter estimates from that model because what basically is happening in that case is the software is forcing some of the parameter to be zero or one, something like that and estimating the other thing, but in this case we are good right now because there are no warning messages, so once you see that convergence criteria met in the output screen here then next thing that you may want to do is go to the log screen and to see if there any warning messages. If there are warning messages don't worry about that model. That model is not good. You may to go to a simpler model than that, but if there are no warning messages then we may want to use these things.

Now let's go for the- to the fixed affects ****. Now if we do a three level design we again see that- we do over here, we'll see that these three things are significant. The classroom intervention is significant. The TV intervention by itself is not significant and the classroom by TV intervention is also not significant, so basically by doing this analysis when we take care of the three levels of nesting it seems like telling the students something in class, do this, do this, when you're seventh graders, eighth graders if we tell them don't smoke, these other bad things that is maybe the most effective way of teaching them about the bad effects of smoking, so in this case now the next thing that we want to look at is this core variance parameter estimates. Again, this will give you an idea about whether something is significant or not, so whenever you want to look at actually in this case if I'm just looking at this one I may just look at this value and this value is fairly big, .993, close to 10%, so it seems like we may not have to worry about adding a school level nesting. The **** level of nesting we already did that. The school, the random intercept saying that all the schools are different from each other, that may not be that important for us that the student when you say random intercept for the school **** at the starting point all the schools are different. That may not be a very important thing that we need to worry about, but definitely the nesting of the class within the school is significant, so if we decide to go to a simpler model again, this is just for an idea. We may again have to do a likelihood ratio test, but in this case just by looking at this if I'm doing the analysis I look at this and say that okay I'm not going to worry about the school level nesting, three levels of nesting. Let me go to a simpler model and just do two levels of nesting, so in that case whether I want to do a class level nesting or a school level nesting I'll be using- instead of using the- instead of having to worry about this model I may be using this model or this model because in this case see I'm estimating one more parameter and in statistics sample size is like money. The more you have the better. You don't want to spend it and in this case if you are estimating a parameter you are basically spending money kind of thing. You are from your data to estimate that particular parameter you are using some of your resources. If that parameter is not relevant, if it's not explaining a lot of things in my data why do I want to spend this thing, resource in this one?

Let me go to a simpler model that is easy to integrate, easy to understand and that does a fairly good job of explaining these things. Now besides that what other thing that I can look at it I may also be looking at how is the parameter estimate changing. Is there a huge change in the parameter estimate by me putting this? If I look at this one there is not that much change in parameter. More than the parameter my interest is again, the standard error because that determines whether something is significant or not, so in this case the standard errors are not changing that much. Here between these two models the standard error is not changing that much also. Parameter estimate is fairly okay, so in that case rather than looking at this model where the school level variance is not significant I may just do a model that has students nested within school or a model where the students are nested within class, so basically what we did is we started with the simplest model, which was the ordinary regression model. We put one level on it with the class, another level with the school and then the final with all three of them, but ultimately we, it seems like, definitely compared to the ordinary regression model these two models that **** of the nesting of the data those are important models, but we may not have to go all the way to the three level nesting where we have schools, classes and the subjects.

Okay, lunchtime. After a long day of biostatistics lunch is always good, so let's eat lunch and then we'll start again.