Key ideas, terms and concepts in SEM

In the first video about structural equation models I gave some backgrounds to what structural equation modeling is the historical path that led to its development some of the key ideas and the ways that it can be applied in social science settings.

In this video I'm going to talk about some of the key ideas terms and concepts in structural equation modeling. This is important because SEM is rather different to other areas of statistics some of the ideas that are important in understanding and applying SEM are quite unfamiliar and so it's important to have a grounding and familiarity with these ideas before we move on to other applications

So in this video I'll be talking about path diagrams the way that we represent equations and theories in the form of diagrams in SEMs. I'll talk about the difference between exogenous and endogenous variables. I'll talk about the way that structural equation modeling analyzes not the raw data but the variance covariance matrix of the variables that were interested in

I 'll talk a little bit about how parameters are estimated using maximum-likelihood in structural equation modeling and I'll also go over how we apply what are called parameter constraints how we don't always estimate every parameters in the model some of the parameters are fixed to values before we start fitting the model. I'll also talk about how we assess the overall fit of a model in structural equation models and the importance of the idea of what it called nested models for assessing model fit and I'll also talk a bit about identification of structural equation models again that's something which is linked to model fit and something that we don't encounter so much in regression context that many people are familiar with.

So the first thing I'm going to talk about path diagrams and path diagram is one of the reasons why structural equation modeling is very appealing to many social scientists in particular. This is because social scientists don't always have such a strong grounding in mathematics and are less comfortable with reading complex equations and so on and so path diagrams are another way of presenting the same information as we can get in a in an equation but they do this visually and that's often a clear way of seeing what is being presented in Equation compared to Greek letters and symbols and so on.

So if we write our path diagrams correctly then we can read directly between an equation and a path diagram they tell us exactly the same things so in this example here we could write a bivariate regression equation in the usual way where our dependent variable Y is a function of our independent variable X and we are going to solve this equation using data and we're going to solve for the unknown parameter beta what is the relationship between X and Y. Now we can we can also write that same information down in the form of path diagram a simple path diagram in this case so we have here Y is now represented as a rectangle X is also a rectangle we have an arrow running from x to y in a single direction and we have a a small circle pointing into y which represents the error term in the equation and you can see that there's a b above the line to indicate that the parameter represented by the straight single arrow is a regression coefficient and this is quite clear I think visually in a sense that what the model is implying at least is that X causes Y and that there is some coefficient beta which summarizes what that causal effect is and there isn't an error in that equation.

There are conventions for path diagrammatic notation so that we use it consistently. There are

some variations in how different conventions are applied and so on but this is the general form where we have a a latent variable is represented in the form of an ellipse an observed variables some variable that we've actually measured in our dataset directly would be represented as in the last slide using a rectangle. Error variances are all small circles and this is similar to a measured latent variable it's a circular form but it's actually a small circle now. This indicates that these error terms are also latent variables but we don't actually label them there is a kind of unknown or residual latent variables. We also indicate the relationships between variables using lines with arrows a curved line with a double arrows at each end indicates a covariance between two variables we call this sometimes a non-directional path or unanalysed relationship because this is used to show that two variables are related to one another but our model does not specify anything about the direction of that relationship it may be because it's not an important part of the theory but we know that the two variables are associated.

Lastly a straight line with a single arrow one end indicates a directional path a regression coefficients so we're saying if we use a single headed arrow then we are indicating the direction of the relationship between two variables in our model. And we can put these basic symbols together to form more complex models but ones which have a clear meaning and which can indeed be translated back into the standard equation notation.

Here are some examples of some quite simple path diagrams here we're just looking at measurement models these are confirmatory factor models and we have here eta1 which is a latent variable shown as an ellipse an eta1 here is shown to cause 3 observed variables X1 to X3 we can also think of that as eta1 the latent variable being measured by the observed variables X1 to X3 and at the top of the diagram we have three error variances e1 to e3 so those are the errors for each of those equations that eta1 is predicting x1 with some error is predicting x2 with error and so on.

So that's a simple path diagram for a factor model and that could be written as an equation but we are in this instance using a path diagram. We can extend this to make a slightly more complicated path diagram now we have two latent variables eta1 and eta2 they are essentially the same diagrams as we saw in the previous slide but now we have two latent variables and we have six observed variable six variables in rectangles each one of which has an error term now we've also added in here a curved line with an arrow at each end this is to show that in our model the two latent variables are correlated with one another we're not saying anything about the direction of the relationship between eta1 and eta2 we're just saying that we think that there is some kind of relationship between them.

In this path diagram we've now introduced a theoretical statement about the direction of the relationship between eta1 and eta2 so we no longer have this curved arrow but we have a straight line with an arrow at one end so what we're saying here is that eta2 is a cause of eta1 and this again would be similar to the first diagram that we saw a bivariate relationship bivariate regression with eta1 regressed on eta2 and we would then have to solve for the unknown beta co-efficient above the straight line with the arrow at the end but this is now a bivariate regression of a latent variable onto another latent variable.

When we are building path diagrams and systems of equations, in structural equation modeling we need to distinguish between two important kinds of variables exogenous variables and endogenous variables. Now an endogenous variable as the name suggests is something which is caused within the system it's a variable that has if you like an arrow

pointing into it it is a dependent variable in one or more equations. And exogenous variable on the other hand is akin to an independent variable in that terminology it's a variable that is not caused by anything within the system of equations that we are presenting as our SEM and that doesn't mean to say that we believe that exogenous variables are in some sense not caused by any other variables it's simply that within our own model the variables in the model it doesn't have any direct calls. Now an important part of SEM is that variables can be both exogenous and endogenous so we can have an arrow pointing into a variable making it an endogenous and that variable itself can have an arrow pointing at another variable making it an exogenous variable in in that limited sense although it's a now a different kind of variable because it has a narrow by pointing into it and an arrow coming out of it and that's important because that kind of variable is a mediating variable it's a variable which through which another variable has an effect on a third variable.

In this path diagram we can which we've already seen but this path diagram now we can distinguish what kinds of variables these are we've got two exogenous latent variables here. They're exogenous latent variables because there is no directional path pointing into either of them. Neither of them therefore has an error term this is just a correlation that we're seeing here so these are both exogenous in the model. Again we've seen this path diagram we've got a new distinction that we can apply to it now though that eta1 is endogenous and eta2 is exogenous. eta2 doesn't have any direct path going into it doesn't have an error term whereas eta1 has an error term pointing into it because it's got a directional path running from eta2.

So a fundamental advantage of using structural equation models is this ability to represent our theories as diagrams rather than using notation which many social scientists are less comfortable with. Another if you like unusual feature of SEM is that in the conventional practice anyway we don't analyze the raw data of the observed variables but we analyze the variance covariance matrix which will denote (S) of those observed variables this is kind of unusual and somewhat surprising I think the people when they first come across it that all the data that we need is just the set of covariances and variances of the observed variables. As we shall see in later videos some structural equation models also use that the means of the observed variables in addition to their variance covariance matrix. So what are we doing with this variance covariance matrix? Well in broad terms we are trying to summarize (S) the various covariance matrix of the observed variables by specifying a simpler underlying structure. So we're going to specify a model which is in some ways simpler than simply reproducing S and our model our SEM in this sense the simpler underlying structure will yield an implied variance covariance matrix what I mean there is that if our model is true then the variance covariance matrix that we observed should look like this it should have these numbers in each of the cells. And again as we'll see later this implied matrix can be compared to the one that we've actually observed and that comparison if it's done properly can tell us something useful about how well our model is accounting for the data to the extent that the implied and the observed matrices differ than our theory i.e. our structural equation model is not doing a very good job of telling us how this data were generated. So a variance covariance matrix probably most people will be more familiar with correlation matrix but here we're dealing with unstandardized variables and this matrix shows six observed variables X1 to X6 and they are in both the columns and in the rows of this table and the diagonal which is shown in bold indicates the variance so the covariance of a variable with itself in this case maybe x1 and x1 that gives us the variance of that variable so covariance of a variance with itself is its variance and those are shown in bold on the main diagonal. Then we see in the other cells the covariances which can be negative or positive in the other cells of this matrix and you'll observe that the top part of the matrix is redundant with the bottom

part so we actually only need the lower part of this matrix.

Now an important aspects of any model fitting, and structural equation modeling is no different, is the needs to estimate what the unknown parameters in our models are i.e. the betas what is the relationship between eta1 and eta2 in the population. Now there are different ways of estimating these parameters in standard regression modeling we would use ordinary least squares. In structural equation modeling practices mainly around using a technique called maximum likelihood and maximum likelihood estimates the unknown model parameters by maximizing the likelihood which we can denote L of a particular sample of data. Now L is the likelihood is a mathematical function which is based on the joint probability of continuous sample observations. So in essence maximum-likelihood finds what the maximum value of L is for a particular sample of data and it does that by sort of iterating through using different values for the unknown parameters until it finds the maximum-likelihood once that maximum has been found then we have produced the maximum likelihood estimates for the unknown parameters. Now maximum-likelihood is appealing because it is unbiased and efficient and now what those terms mean are that if we have a large sample then our estimates of the unknown parameters will be correct they will converge upon the true values in the population. They are efficient in the sense that no other way of doing this will give us more precise estimates of those parameters. Now those two assumptions of being unbiased and efficient do themselves hinge on some other assumptions one important one is that the data come from a multivariate normal distribution essentially that requires us to be using continuous variables so maximum likelihood is less good when we have variables in our datasets that are not continuous and that we have arrows pointing into. In those situations we need to use different estimators but for now I'll be focusing on the simpler case of multivariate normal data and maximum-likelihood.

Now another way in which maximum likelihood is used in SEM is that not just in the estimation of the unknown parameters but in use of the likelihood the maximum likelihood. If we take the log of the likelihood for the model then we can use this to test how well our model fits compared to some more or less restrictive alternative so maximum likelihood is used in in two ways in SEM one is for estimation of the unknown parameters and linked to that is use of the log likelihood to assess how well the model fits the observed data.

Most areas of statistics that social scientists are familiar with the focus is very much on estimating the unknown parameters in the model we want to know what the relationship between X and Y is in the population or possibly what the conditional association between two variables is and so we we focus on estimating those unknown parameters It's also true of course in SEM but in SEM we have an additional focus which is on fixing or constraining parameters to particular values before we estimate our model and that's a bit unusual for many people. So we can fix model parameters to any values but it tends to be the case that we will be fixing parameters to be the value 0 or the value 1 those are the most common parameter constraints that we make in SEM and I 'll come back to why we do that later. But we can also in addition to fixing parameters to these values we can also constrain model parameters to be equal to other model parameters so we will still estimate those equivalised parameters but they have to be estimated so that they are the same the model applies that constraint on what the parameters are that are estimated so again that's something which is quite unusual and we don't really see that in many other statistical techniques that we might use in social science. The main thing that we are using these parameters constraints for is for the purposes of model identification and I would be saying some more about that so. Now I said that we can use the likelihood of our model to test how well it fits the data by comparing

L model with another model now when we do this the two models that we compare have to be what is called nested one within the other. So what do we mean by nested? Well it is precisely that one model is a subset of the other or the parameters in one model or a subset of the parameters in the other model. Another way of saying this is that if we have two models A and B then model A is the same as model B but it just adds some additional parameter restrictions so A is B + parameter restrictions.

To take an example here then if model B has the form Y= a + b1X1 + b2X2 +e then model A will be nested if it has that same structure but it applies a parameter constraint to the two beta co-efficient that they are equal. So we now have this property that model A is the same as model B with an added parameters constraint it is therefore nested within model B. if we consider a third model C though and we now remove X2 from the model and we add Z2 instead then model C is not nested within model B because it isn't just model B+ the parameter restrictions it has a new variables Z which is not in model B. So these are if you like apple and pear models we can't really in any sensible way make comparisons between the fit of model B and of model C because they include different variables.

So I said something about model fit already and the fact that it is based on the log of the likelihood of the model that we've estimated and that we can do this comparison of model fit when the two models are nested this is because if we take the log of the likelihood for model A from the likelihood for model B the log of the likelihood then that is itself that number is itself distributed as Chi square and the Chi square distribution then has a degree of freedom which is equal to the difference in the degrees of freedom for Model A and Model B. We can therefore use this Chi square distribution to test the fit of the first model to the second model. Now if our value of Chi square has a P value greater than 0.05 then we will prefer the more parsimonious model A because what we're saying here in this situation is that the models are not different with regard to one another's likelihood values that we say that the likelihoods are essentially the same and that means that we will prefer the model that has the simpler and is estimating fewer parameters. So we're in this case the model B would be our observed data, the variance covariance matrix, then we're saying that there is no difference between the observed and the employed matrices and our model therefore fits the data well so that's the essence of the assessment of model fit using Chi square in structural equation models we can look at the difference in the likelihood for one model and compare it to the likelihood for a nested model and make a statistical test of whether one fits the data better than the other.

So the last thing I'm going to talk about in this video is model identification this is all linked with the things I've talked about already to parameter constraints and fixing parameters to particular values into assessing model fit and so on. So what is model identification? Well in conceptual terms we need to have enough known pieces of information in an equation to produce unique estimates of unknown parameters and we need unique estimates otherwise we don't know which ones to prefer. So to give an example of what we made here by the balance between known and unknown pieces of information if we look at these two equations the first of these is unidentified we have X + 2Y =7 So what we would want to do is to find the unique value that satisfies that equation for y . Now because we have a balance of knowns and unknowns where X and Y could really take on many many different values and they would all be true if you like in terms of the equation being correct that equation is unidentified because it doesn't enable us to produce unique estimates now if we change that equation slightly where we now are not having X an unknown and we make that 3 then we can only have one value for Y which is 2 in a way that will satisfy that equation so that equation then is identified. Now that is the essence of what we need to understand about

identification is that it's to do with the balance between the number of known and unknown pieces of information in an equation. Now there is something else to know about identification which is that it says a theoretical property of the model it's not really linked to the data as such so we can figure out what the identification status of a particular model is without having any data or estimating any parameters but it's also true to say that a model can be theoretically identified but empirically unidentified given a particular set of data. So we are looking at the balance between the known and the unknown piece of information in our equations and in SEM the known pieces of information are the variances and covariances in means if we're using means in our model of the observed variables these are the known pieces of information. The unknown pieces of information are the parameters that we want to estimate in the model.

Now models can have different identification status. A model can have as we saw a moment ago more unknowns than knowns that means it is unidentified we can't produce unique values for the unknown parameters that's an unidentified model. Other models can be just identified where the number of knowns is equal to the number of unknowns we don't have any what we call over identifying restrictions on the model and therefore for just identified models we don't have any likelihood for the model that we can use to assess its fit. Now most of the models that people are familiar with using again ordinary least squares regression those kinds of models are just identified.

The third level of identification status is over identified models and that's usually what we are trying to get to and deal with in SEM and that's where the number of knowns is greater than the number of unknown parameters in the model and that means that we can assess the fit of the model as well as estimating the unknown parameters.

So there are different ways that we can assess the identification status of a model a very simple one these days with modern computers is simply to run our model and most software will tell us what the identification status is of the model even before we fit the data so it's quite easy compared to how things were done in the past but nonetheless it is still useful to have a consideration of the identification status of a model as it helps us to understand where things might be going wrong if we have a problem and our model is unidentified working through it in this way can help us to see why. So here is the accounting rule that can be used where if we have S= number of observed variables in the model and the number of observed variables in the model and the number of non-redundant parameters is given by this equations i.e. 1/2 of S times (s +1) again s the number of observed variables and t is the number of parameters that we are going to estimate in the model number of unknown parameters so if t is greater than the answer to this equation then our model is unidentified we have more unknown parameters than we have non-redundant parameters and if it's less then we have an over identified model.

So to give an example of that here is the path diagram that we saw earlier where we have e1 a latent variable which is measured by or causing 3 observed variables and each of those observed variables has an error variance. So if we want to find the number of non-redundant parameters we can use our 1/2s (s+1) equation then we have s here is equal to 3 so s (s+1) is 3(4) that's 12 if we take half of that that gives us 6 as the number of non-redundant parameters. Now how many parameters we are trying to estimate with this model? Well 3 variances one for each error term we've got 2 factor loadings one of them you'll see there is constrained to 1 so we're fixing that loading and that is for identification of the models so we're not estimating that factor loading but we are estimating the other two so we have 2

factor loadings and then lastly we have variance for the latent variable so 3+2 +1 is 6 parameters to be estimated which is the same as the number of non-redundant parameters so we have with this model zero degrees of freedom the model is just identified. So we can estimate the unknown parameters but we do not have any way of assessing the fit of this model because it's just identified no degrees of freedom.

Now something else that's important to understand about identification is that we as the analyst can control to some degree the identification status of our model so we can do this for a model like the one that we just saw that's just identified or model that is under identified by adding more known pieces of information to the equation or by removing some unknown pieces of information removing parameters that are to be estimated and adding constraints. So if we were to constrain two of the parameters in the model to be equal to one another let's say we constrain two of the regression coefficients of the factor loadings to be equal now we're only estimating one parameter where previously we were estimating two, so we've removed one unknown and gained one degree of freedom. Now we can see this in this model here where we have added an additional observed variable to the previous path diagram so now the model is essentially the same but we've got a fourth observed variable X4. We now are estimating an additional factor loading and an additional error variance but we have gained more in terms of our known parameters so now if we use our 1/2s (s+1) equation, s is now 4 so 4 (s+1) becomes 20, we take half of that now we have 10 non-redundant parameters in this model and we have 4 + 3 + 1 parameters to be estimated 8 so 10-8 gives us 2 degrees of freedom so by adding that fourth observed variable our model is now over identified and we can say something about the fit of that model to the variance covariance matrix that we observed.

So in that example we changed the identification status of our model by adding in more known another known piece of information and other observed variable. Another way of changing identification status is to remove unknown parameters. Now here in this example we are now not estimating the two factor loadings that we were in the first example so you can see there's a number 1 next to each of the arrows for the factor loadings so rather than estimating those we're saying these are all equal this may not be a very theoretically meaningful thing to do this isn't the point at this particular juncture what we're showing here is that you can change the identification status of the model by removing unknown so we're not estimating these anymore so we still have six non-redundant parameters but we now are only estimating 4 unknown parameters because we are not estimating any of the factor loadings so now this model is over identified.

So in this video I have covered some of the important ideas and concepts that learners will need to take with them into later videos and applications. These are focused around the use of path diagrams for representing our theories and their equations. The fact that we analyzed the variance covariance matrix of the observed variables rather than the raw data, which we use for the most part maximum likelihood estimation which has quite restrictive assumptions about multivariate normality but nonetheless is a very useful estimator. It gives us consistent unbiased and efficient estimates of the unknown model parameters and allows us to do global tests of the fit of the model to our data. Those kinds of fit tests are mainly applicable in the context where models are nested where we can say that one model is a subset of a second model that it is the same as the second model with some additional parameter restrictions. And I talked about identification of models and models that can be under identified just identified and over identified and how we as the analysts can exert some control over the

identification status about our model by removing unknown parameters or adding in more known parameters.