Structural equation models (also referred to as "SEM models") have become very popular in the Social Sciences, especially in Psychology, Sociology, Education and, more recently, in Business and Public Administration and various applied health sciences (e.g., Nursing). A major feature in the development of structural equation models from the earlier causal ("path") models of the 1960s and 1970s is the conceptualization of latent variables. The terms, "unmeasured variable models" and "latent variable models" refer to types of structural equation models that explicitly incorporate measurement error into the estimation of structural equation parameters, and treat observed ("manifest") variables as indicators of underlying constructs rather than perfectly measured representations of these same constructs. These models are quite general, and subsume many of the multivariate techniques typically dealt with in lower-level courses, including regression models, factor analysis, analysis of variance/analysis of covariance, principal components analysis, path modeling, mediation analysis and moderation analysis. More recently, SEM models have provided an approach to the estimation of parameters in growth curve models for longitudinal data (these can also be estimated in the multilevel model framework), and for an approach to the problem of the unbiased estimation of parameters in the presence of missing data. Additional recent extensions have generalized models from those which required measured variables to be continuously distributed to those suitable for categorical data as well as the application of multiple-level models (fixed- and random-effects models) to the SEM framework.

In earlier usage, the models discussed in this course were, in the 1980s and early 1990s, often referred to as LISREL models. At the time, almost all SEM research used the computer software called LISREL. Today, SEM models can be estimated using almost any statistics package as well as through stand-alone programs, the most popular and comprehensive of which is MPlus. SEM programs are now available built in to STATA and SAS, and as an add-ins and SPSS and as R packages.

In this course, we will emphasize STATA and R for most models. The course, including course assignments for those taking the course for credit, is structured in such a way that participants can choose either R or STATA, although some participants may elect to work with both. For some more advanced models, we will be presenting models with MPlus software.

The course starts with an introduction to single-indicator causal models involving intervening variables (mediators), and then progresses into models with multiple indicators for some or all of the constructs. After this introduction to SEM models in scalar terms, and an introduction to the extension of our models to simultaneous multiple-group estimation, we will briefly introduce the matrix-form representation of SEM models on an “if time permits” basis. Much of the early literature presents models using LISREL matrix notation (though this is no longer the norm), and some software (e.g., LISREL) is programmed primarily in matrix form. Next, we extend the models we have learned to models for means and intercepts, which are useful both for factor mean comparisons in the multiple-group case and for longitudinal data analysis in the single (or
multiple) group case. Towards the end of the course, we shall cover some more advanced topics, including estimation in the presence of missing data and growth curve models for longitudinal data. These topics both require a thorough understanding of models for means and intercepts, which are usually covered in week 3. As time permits, we will cover extensions to models with ordinal or binary indicators but continuous latent variables, and to models with categorical latent variables (also known as latent class models); in recent years, we have usually been able to cover these topics, at least in a preliminary fashion, in week 4.

What sort of a background is required for this course? A thorough familiarity with regression models is absolutely essential. At the very least, individuals should have taken the I.C.P.S.R. Regression Analysis II workshop or its equivalent (note that this is a second level graduate regression course; a basic “introduction to regression” course is probably not sufficient). An understanding of the rudiments of matrix algebra will be needed for a small part of the class (the discussion of SEM models in matrix notation, usually taking up a single class). I.C.P.S.R. offers a set of Matrix Algebra Lectures early in the second session; participants who have little exposure to matrix algebra might consider attending these lectures. Individuals taking the course for formal credit or requiring a “grade letter” are warned that the absence of at least some introduction to matrix algebra could prove to be an impediment to achieving excellent grades. Some exposure to factor analysis will be helpful, since there are distinct parallels between some aspects of SEM modeling and factor analysis, but should not be considered essential. Finally, for the categorical data extensions discussed in week 4, some prior exposure to models for categorical dependent variables (typically discussed in Regression II and discussed in greater detail in the Categorical Data Analysis) will be helpful.

**Required and Recommended Readings:**

The major readings for this course can be found in a series of 15-30 page written course notes which will be posted on server locations to be identified to ICPSR program participants. These are exclusively for the use of ICPSR participants and are in PDF format, and can be downloaded.

While the course notes cover most of the material dealt with in the course, participants may wish to purchase copies of an additional text, since the ability to “triangulate” explanations is sometimes helpful in learning new techniques. The most useful of these will be:


A bit more advanced:


Note, though, that the course notes supplied online should be sufficient; purchasing either or both of the 2 abovementioned texts should not be considered essential. Also, these textbooks tend not to follow the same sequence of presentation as the course.

The following texts cover some of the more advanced topics we deal with.


An important comprehensive resource for basic and advanced topics can be found in the edited Handbook of Structural Equation Modeling (Guilford Press, 2012).

Software

Access to STATA software can be obtained via the “virtual machines” provided by the summer program. Check to see if there is a special summer program STATA license which allows you to download STATA and use it on your laptop for the duration of the program. If you want a permanent copy of STATA, there may be a summer program “Grad Plan” (special price STATA purchasing plan for graduate students), though you may be able to get this sort of pricing via your home institution as well. STATA comes in different “versions” all of which offer the full suite of software procedures. The versions differ mainly in the number of variables that can be handled and to some extent in terms of speed. The cheapest “IC” version, which can handle files of up to 2,000 variables (and 2 billion cases) is all that is required for the course. The more expensive SE version only adds the ability to handle files with more variables (and huge models with more than 800 independent variables – not an issue with this course), while the even more expensive MP version is a bit faster on multiple-core computers.

For those wishing to develop an ability to work with SEM models in R, access to R software is straightforward as well: you will want to download and install R from a CRAN mirror site, along with the haven package and the lavaan package. The semTools package will also be helpful.

MPlus software for advanced SEM modelling is extremely expensive. The free trial version is not worth the effort one would employ to install it; it is too restrictive even for learning purposes. MPlus software will be available on the virtual machines participants can log into. Given the price of MPlus, most participants may wish to hold off on purchasing the software until after they have finished the course (and decided, on the basis of what they have learned, that they really need to get MPlus as opposed to being able to work with SEM procedures in STATA or R). If MPlus is purchased, the most expensive version – the “Combination Add-in” – should be obtained (the base program does not, for the most part, do anything that cannot be done with STATA, R or other SEM software). Beware of purchasing the student version as it cannot be renewed beyond 3 years.

Assignments and Exercises

Most participants in this workshop do not attend for the purposes of obtaining formal course credit. For non-credit participants, it is important to complete as many of the computer exercises as possible; without practical experience working with software and writing up
“results,” participants are not likely to be able to conduct research of their own using the methods discussed in the course. There will be five exercises (one “pencil and paper” and 4 computer exercises) throughout the course; these are submitted and returned with comments (and, in the case of credit participants, a grade). For credit participants, there will also be two take-home tests. The first of these is “required,” while the second test can be skipped if all other assignments are completed unless one is attempting to obtain a grade of “A+”.

It is important that individuals who require a grade at the end of the course (taking the course for formal credit or would like ICPSR to write a letter indicating the grade that was received) identify themselves at the beginning of the course or mark “credit” or “grade” on their assignments. Assignments submitted by non-credit participants not requiring a grade are returned with comments and suggestions, but not with a grade.

Participants requiring a grade should ask for a copy of a “Grade Information” sheet which provides further information on the computation of formal grades for the course.

In the past, participants have asked if it would be possible to substitute any course requirements for a “major project” involving data that they are interested in working on. Unfortunately, the brevity of the summer program makes this alternative form impossible.

**Topic Outline:**

*The relationship between sub headings and the daily lecture schedule is approximate, and will vary according to the amount of time devoted to particular topics (which itself varies from class to class given variables such as the level of background of class participants, questions asked in class, etc.). Readings do not always perfectly cover topics discussed in lectures (a # signifies that the match is particularly weak for a given reading); additional details can be found in the powerpoint slides posted for the class. Please note that the page/chapter references to Schumaker and Lomax are for the 3rd (2010) edition, which is available at the ICPSR library, not the 4th (2015) edition, which may or may not be available.*

**Week 1**

a) Basics: path analysis principles important in SEM modeling; manifest vs. latent variables; measurement error and its implications. Working with covariances. Limitations; relationship to OLS assumptions. Special cases and extensions. Software available for the estimation of SEM models. *Reading: Class manuscript chapter 1.* #Reading: Hoyle, chapter 1; #Schumacker and Lomax, chapter 3.

b) Covariance algebra for SEM models; Hypothesis testing (testable and non-testable hypotheses); examples using STATA and lavaan

*Reading: Class manuscript, chapter 2*

c) Estimation; interpreting results. Path model examples. Identification (the basics)

*Reading: Class manuscript, chapter 3; #Schumacker and Lomax, chapter 4 (a bit more advanced, Kenny and Milan, chapter 9 in Hoyle).*

d) More on identification; Scaling and interpretation issues; constraints in SEM models; variances of latent variables; model fit and model improvement; dummy exogenous
variables  Reading: Class manuscript, chapter 4; Hoyle chapter 5; Schumacker and Lomax, chapter 5

e) Equality constraints; extended presentation of a latent variable model example; problems and issues (negative error variances, non-convergence): item parcels (pro and con).
Reading: class manuscript, chapter 5; #chapter 7 of Schumacker and Lomax for review and extensions.

Week 2

a) Multiple-group models (the 2-group case); measurement model equivalence across groups programming multi-group models in STATA Class manuscript, chapter 8
b) Multiple group models (the 3+ group case); construct equation equivalence; parallel and non-parallel slopes; Wald, LM and score tests in multiple-group models; interpretation of modification indices higher order factor models; programming SEM in R
Reading: Kaplan, chapter 4; Schumacker and Lomax, chapter 13 (sections 13.1 and 13.2 only).
c) SEM assumptions; SEM problems (more detail): inadmissible solutions; convergence issues; local minima and solutions (more detail): user provided start values, model respecification. More on multiple group models: a four-group example. Testing blocks of variables in multiple-group models. Reading: #Schumacker and Lomax, chapter 11; Hoyle and Panter, chapter 9 in Hoyle.
e) Missing data. Reasons not to use listwise or pairwise deletion. Reasons why single-regression imputation and mean substitution are simply terrible. Multiple-group modeling for missing data; EM covariance estimation; multiple imputation; full information ML estimation; nearest neighbour; auxiliary variables in MI and FIML approaches.

Week 3

a) Indicator and effects coding for variance scaling. Means and intercepts in SEM models. Application to longitudinal data; Application of mean/intercept models to multi-group problems. Factor mean comparisons Programming considerations. Reading: Course reading notes chapter 7; Schumacker and Lomax, 13.3, 13.4
b) Exercises and examples (mean and intercept comparisons) Parallel and non-parallel slope models. Graphing outcomes in non-parallel slope models.
c) Parallel and non-parallel slopes in 3+ group models.
d) Non-normally distributed indicators in SEM models. The ADF estimator. Scaled test statistics. Bootstrapping (note: this presentation is within the general confines of models which assume that the underlying latent variables are continuous and that the manifest variables are at worse
coarsely categorized continuous variables; see week 4 for models for categorical latent variables). Reading: Edwards et al., chapter 12 in Hoyle.

Week 4

a) Growth curve models for multiple-indicator variables. Curve of factors and factor of curves models Reading: #Schumacker and Lomax, 16.2. Other readings TBA
c) Mixture models for categorical outcomes Reading: #Schumaker and Lomax, 15.2; #Kaplan, chapter 9.