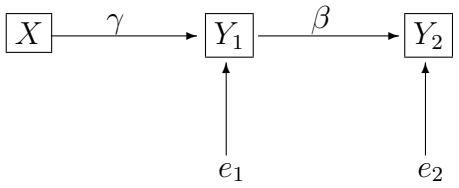
STA 431s11 Assignment 9

Do this assignment in preparation for the quiz on Friday, March 18th. Answers to the non-SAS questions are practice for the the quiz, and are not to be handed in. For Question 5, bring both your log file and your list file to the quiz; they may (or may not) be handed in.

1. Consider this model:



   (a) Express the model as a set of equations. Please start by writing "Independently for $i = 1, \ldots, n, \ldots$" and put a subscript i on all the random variables. Assume that everything is normal with mean zero, and give the variances and covariances of all the exogenous variables in the model (including error terms). Make up your own symbols for parameters when necessary.

   (b) What is the parameter θ? It has 5 elements.

   (c) Is this model saturated? Answer Yes or No.

   (d) Classify the random variables as either Exogenous or Endogenous, and as either Manifest or Latent.

   (e) What is the joint distribution of the manifest variables? Express the variance-covariance matrix in terms of the parameter θ; show your work. Each cell in your matrix should contain a formula for the variance or covariance in terms of quantities like ϕ, β and so on.

   (f) Are the parameters of the model identifiable? Answer Yes or No and prove it.

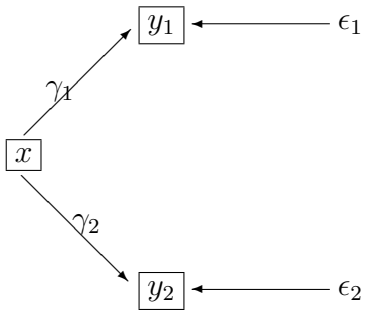
2. The following model has zero covariance between all pairs of exogenous variables.

$$\begin{aligned} Y_1 &= \gamma_1 X + \epsilon_1 \\ Y_2 &= \beta Y_1 + \gamma_2 X + \epsilon_2 \\ W &= X + e_1 \\ V_1 &= Y_1 + e_2 \\ V_2 &= Y_2 + e_3 \end{aligned}$$

   (a) Draw the path diagram. Put a coefficient on each straight arrow, either the number one or a Greek letter.

   (b) As the notation indicates, the observable variables are W, V_1 and V_2. Are the parameters of this model identifiable at every point in the parameter space? Respond Yes or No and justify your answer.

3. Consider the following model:



   (a) Write this model as a set of simultaneous equations. Please start by writing "Independently for $i = 1, \ldots, n, \ldots$" and put a subscript i on all the random variables.

   (b) Give symbols for all the non-zero variances and covariances for the exogenous variables in the model. Please use the standard notation.

   (c) What is the parameter vector $\boldsymbol{\theta}$ for this model? It has five elements.

   (d) Since this is a regression model, the parameters are identifiable. Are they *just* identifiable? Is the model saturated? Answer Yes or No to both questions.

   (e) In a test of model fit, what would the degrees of freedom be? The answer is a single number.

4. In the *double measurement design*, two parallel sets of measurements are taken on a collection of latent variables (endogenous, exogenous or some of each). The two sets of measurements are usually collected on different occasions and in two different ways. Great pains are taken to ensure that the measurement errors are uncorrelated between sets; correlation *within* each set is allowed. For example, the cases could be companies, with one set of measurements consisting of information provided by management. If the same information were independently determined by a forensic audit, it might be safe to assume that the measurement errors were uncorrelated across occasions, though errors of measurement within sets (especially by the company) would almost certainly be correlated.

   This not very interesting as a factor analysis model, because all the factor loadings equal one. But it is useful as a measurement model for more general structural equation models with latent variables. In the following, all the latent variables are collected into a "factor" $\mathbf{F}$, and the observable variables are collected into $\mathbf{D}_1$ (measured by method one), and $\mathbf{D}_2$ (measured by method two). Let

$$\begin{aligned} \mathbf{D}_1 &= \mathbf{F} + \mathbf{e}_1 \\ \mathbf{D}_2 &= \mathbf{F} + \mathbf{e}_2, \end{aligned}$$

   where all expected values are zero as usual, $V(\mathbf{F}) = \boldsymbol{\Phi}$, $V(\mathbf{e}_1) = \boldsymbol{\Omega}_1$ and $V(\mathbf{e}_2) = \boldsymbol{\Omega}_2$; all the covariance matrices are positive definite. The random vectors $\mathbf{F}$, $\mathbf{e}_1$ and $\mathbf{e}_2$ (all k by 1) are independent. Notice that the variance-covariance matrices of the error terms need *not* be diagonal. This is a big improvement over a model in which *all* the measurement errors are assumed to be independent.

   (a) Write $\boldsymbol{\Sigma}$ as a 2×2 partitioned matrix (a matrix of matrices).

   (b) Prove that the parameter matrices $\boldsymbol{\Phi}, \boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are identifiable.

   (c) Assuming no further constraints on the parameters, how many equality restrictions are there? Your answer is a single number that depends on k.

5. The file `littlemath.data` (see link on the course web page in case the one from this document does not work) contains data from a study of students taking first-year calculus on one campus of a large North American university. The students took a diagnostic test with two parts: pre-calculus and calculus. Marks in high school calculus were available, as were marks in first-year university calculus. The question is whether these data are unidimensional. That is, do data arise from a single underlying factor ("Math Skill" or something)?

   Recall that while the factor loadings from an exploratory factor analysis are usually worthless, the test for number of factors is pretty good. So, use `proc factor` to check for unidimensionality (because it's a lot easier than `proc calis`). What you are seeking is two numbers: a test statistic and a p-value. Know what each one means. Are the data unidimensional? Answer Yes or No.