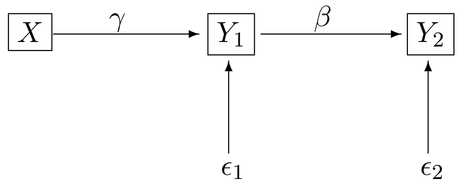
STA 431s17 Assignment Nine[1]

The non-computer questions on this assignment are practice for the quiz, and will not be handed in. Please bring your log files and your results files for the SAS part of this assignment (Question 10) to the quiz. There may be one or more questions about them, and you may be asked to hand printouts in with the quiz.

1. In the following model, assume that $E(X) = \mu_x$, and the regression equations *do* have intercepts.



   (a) Classify all the random variables in the model (including error terms) as either Exogenous or Endogenous, and as either Observable or Latent.

   (b) 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 all the exogenous variables are normal, and include this in the statement of the model. Make up your own symbols for parameters when necessary, but try to stay consistent with the notation being used in the course.

   (c) What is the parameter vector $\boldsymbol{\theta}$ for this model?

   (d) What is the joint distribution of the observable variables? Express the mean vector and variance-covariance matrix in terms of the model parameters; show your work. Each element of the mean vector and the variance-covariance matrix should contain a formula in terms of quantities like ϕ, β and so on.

   (e) Are the parameters of the model identifiable? Answer Yes or No and prove it. If the answer is No, all you need is a simple numerical example of two distinct parameter vectors that yield the same mean and covariance matrix of the observable data.

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

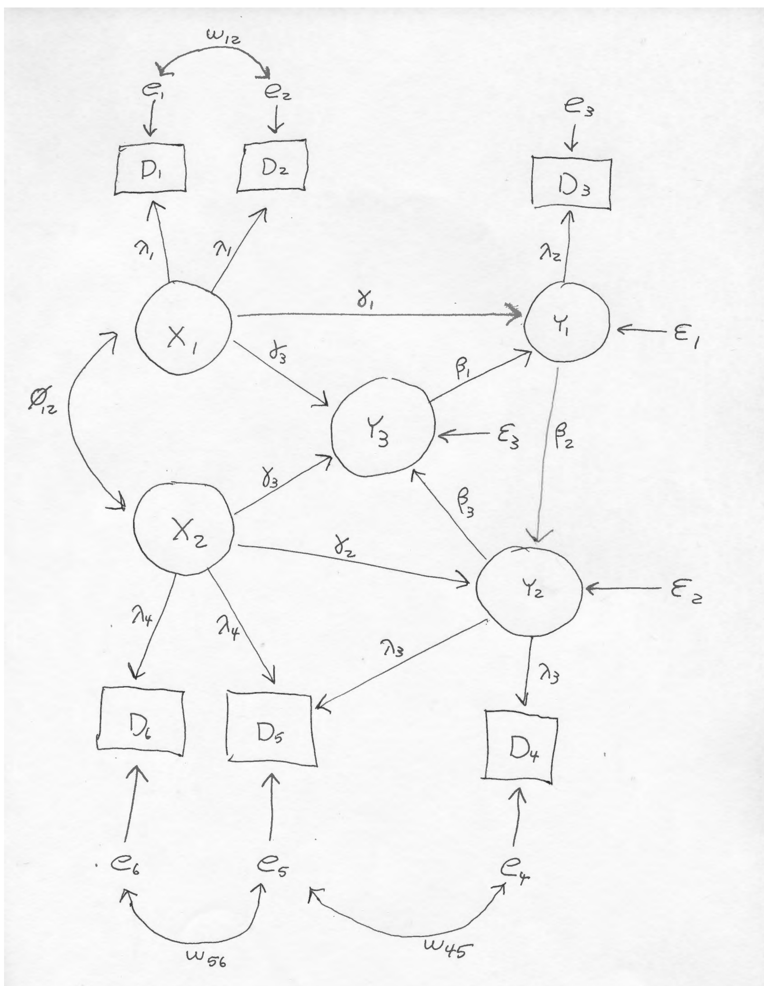
   (g) Suppose that X, Y_1 and Y_2 were all latent variables, and there were two independent measurements of each one. Independent means no covariance between any measurement errors. For simplicity, assume no intercepts and $E(X) = 0$.

      i. Draw the path diagram for the new model.

      ii. Are all the parameters of the new model identifiable? Answer yes or no and explain why. No detailed calculations are needed.

2. Patients with high blood pressure are randomly assigned to different dosages of a blood pressure medication. There are lots of different dosages, so dosage may be treated as a continuous variable. Because the exact dosage is known, this variable is observed without error. After one month of taking the medication, the level of the drug in the patient's bloodstream is measured once (with error, of course), by an independent lab. Then, three independent measurements of the patient's blood pressure are taken. One is done by the lab that did the blood test, one is the average of 7 daily measurements taken at home by the patient, and one is done in the doctor's office. Notice that the same lab measures the blood level of the drug, and also does one of the blood pressure measurements. Do *not* assume that errors in the two measurements carried out by the lab are independent. **Make a path diagram. Do not bother to write coefficients on the arrows this time**, but write brief labels ("Dose" etc.) in the boxes and ovals.

3. Consider the following path diagram.



(a) Give the model equations in scalar form.

(b) Give the latent variable model equations in matrix form: $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$. Use symbols from the path diagram. For example, you will write $\boldsymbol{\beta}$ as a 3×3 matrix of zeros and β_j symbols.

(c) Give the measurement model equations in matrix form: $\mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i$. Use symbols from the path diagram.

(d) Give the following matrices: $\boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}$. Make sure the dimensions are correct. Some of the symbols you need are on the path diagram, but not all.

4. For the General Structural Equation Model (see formula sheet), calculate

   (a) $cov(\mathbf{Y}_i)$

   (b) $cov(\mathbf{X}_i, \mathbf{Y}_i)$

5. In your calculation of $cov(\mathbf{Y}_i)$ and $cov(\mathbf{X}_i, \mathbf{Y}_i)$, you used the matrix $(\mathbf{I} - \boldsymbol{\beta})^{-1}$. As described in lecture, the existence of this matrix is implied by the model. Assume it does *not* exist. Then the rows of $(\mathbf{I} - \boldsymbol{\beta})$ are linearly dependent, and there is a $q \times 1$ vector $\mathbf{v} \neq \mathbf{0}$ with $\mathbf{v}^\top(\mathbf{I} - \boldsymbol{\beta}) = \mathbf{0}$.

   (a) Under this assumption, show $\mathbf{v}^\top \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^\top \mathbf{v} = -\mathbf{v}^\top \boldsymbol{\Gamma} \mathbf{X}_i \boldsymbol{\epsilon}_i^\top \mathbf{v}$.

   (b) Show that this contradicts $\boldsymbol{\Psi}$ positive definite.

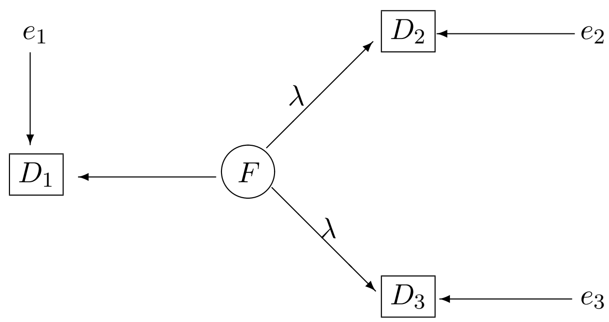
6. The following centered model has zero covariance between all pairs of exogenous variables, including error terms.

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

   (a) Draw the path diagram. Put a coefficient on each straight arrow that does not come from an error term, either the number one or a Greek letter. It is assumed that all straight arrows coming from error terms have a one.

   (b) As the notation suggests, the observable variables are W, V_1 and V_2. Are the parameters of this model identifiable from the covariance matrix? Respond Yes or No and justify your answer.

7. Consider the following model.



(a) Write the model equations in centered form. Please start by writing "Independently for $i = 1, \ldots, n$, ..." and put a subscript i on all the random variables.

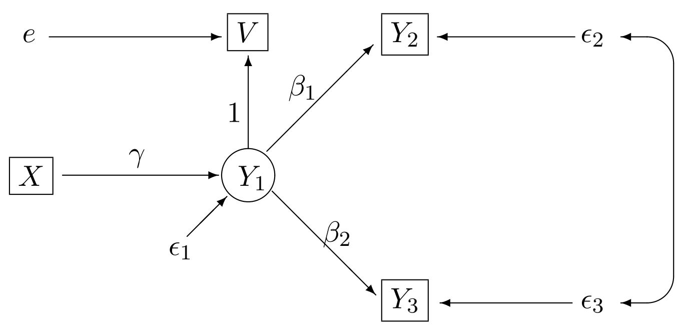
(b) Let $\boldsymbol{\theta}$ denote the vector of parameters that appear in the covariance matrix of the observable data. What is $\boldsymbol{\theta}$?

(c) Does this model pass the test of the parameter count rule? Answer Yes or No and give the numbers.

(d) Are the elements of $\boldsymbol{\theta}$ identifiable from the covariance matrix? Answer Yes or No and prove it. If the answer is No, all you need is a simple numerical example of two distinct parameter vectors that yield the same covariance matrix of the observable data.

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

8. In the following model, all random variables are normally distributed with expected value zero, and there are no intercepts.



(a) Write the model equations in scalar form.

(b) What is the parameter vector $\boldsymbol{\theta}$ for this model? Use standard notation. Include unknown parameters only.

(c) Does this model pass the test of the parameter count rule? Answer Yes or No and give both numbers.

9. Make a path diagram for the following data set. A farm co-operative (co-op) is an association of farmers. The co-op can buy fertilizer and other suppies in large quantities for a lower price, it often provides a common storage location for harvested crops, and it arranges sale of farm products in large quantities to grocery store chains and other food suppliers. Farm co-ops usually have professional managers, and some do a better job than others.

   We have data from a study of farm co-op managers. The response variable of interest is job performance, a latent variable. The variables in the "latent variable" part of the model are the following, but note that one of them is assumed observable.

   X_1: Knowledge of business principles and products (economics, fertilizers and chemicals). This is a latent variable measured by W_{11} and W_{12}.

   X_2: Profit-loss orientation ("Tendency to rationally evaluate means to an economic end"). This is a latent variable measured by W_{21} and W_{22}.

   X_3: Job satisfaction. This is a latent variable measured by W_{31} and W_{32}.

   X_4: Formal Education = Number of years of formal schooling divided by 6. This is an observable variable, assumed to be measured without error.

   Y: Job performance. This is a latent variable measured by V_1 and V_2.

   The data file has these observable variables in addition to an identification code for the managers.

   W_{11}: Knowledge measurement 1

   W_{12}: Knowledge measurement 2

   W_{21}: Profit-Loss Orientation 1

   W_{22}: Profit-Loss Orientation 2

   W_{31}: Job Satisfaction 1

   W_{32}: Job Satisfaction 2

   X_4: Formal education, assumed measured without error

   V_1: Job Performance 1

   V_2: Job Performance 2

   In this study, the double measurements are obtained by just splitting questionnaires in two, as in split half reliability. Furthermore, all the measurement errors are assumed independent of one another. This is consistent with mainstream psychometric theory, though maybe not with common sense. For this assignment, please assume that the errors are independent of one another, and independent of the explanatory variables. The explanatory variables, of course, should *not* be assumed independent of one another.

10. The file manager.data.txt has raw data for the study described in Question 9. This is a reconstructed data set based on a covariance matrix in Jorekog (1978, p. 465). Joreskog got it from Warren, White and Fuller (1974).

   (a) Using proc calis, fit the appropriate model. There are 98 co-ops, so please make sure you are reading the correct number of cases. For comparison, my value of Akaike's Information Criterion (which will not be on the quiz) is 73.1819. If you get this number, we must be fitting the same model. Using your results file when necessary, be ready to answer questions like the following on the quiz.

      i. There is one observable exogenous variable. What is it?
      ii. There is one latent endogenous variable. What is it?
      iii. Based on the number of covariance structure equations and the number of unknown parameters, how many equality restrictions should the model impose on the covariance matrix? The answer is a single number; you need not say exactly what the equality restrictions are.
      iv. Does your model fit the data adequately? Answer Yes or No and give three numbers: a chisquare statistic, the degrees of freedom, and a p-value.
      v. Controlling for knowledge, profit-loss orientation and job satisfaction, is there evidence that formal education is related to job performance? Answer Yes or No and give the value of a test statistic (actually it's a Z) that supports your conclusion. Of course in all these questions you are using the $\alpha = 0.05$ significance level and a 2-sided test.
      vi. Controlling for formal education, knowledge and profit-loss orientation, is there evidence that job satisfaction is related to job performance? Answer Yes or No and give the value of a test statistic (actually it's a Z) that supports your conclusion. If the answer is Yes, say whether satisfaction is positively related to performance, or negatively related.
      vii. Controlling for job satisfaction, formal education and knowledge, is there evidence that profit-loss orientation is related to job performance? Answer Yes or No and give the value of a test statistic (actually it's a Z) that supports your conclusion. If the answer is Yes, say whether profit-loss orientation is positively related to performance, or negatively related.
      viii. Carry out a Wald test of all the regression coefficients at once; use the `simtests` command. Be able to give the value of the chi-squared test statistic, the degrees of freedom, and the p-value – all numbers from your printout. Using the usual $\alpha = 0.05$ significance level, is there evidence that at least one regression coefficient must be non-zero? As a by-product, you get four tests with p-values identical to some in the default output. What null hypotheses are they testing?
      ix. Estimate the reliability of Knowledge measurement 1. Your answer is a number. You will need a calculator.
      x. Estimate the reliability of Knowledge measurement 2. Your answer is a number. You will need a calculator.

xi. Show that the reliabilities of Knowledge Measurements 1 and 2 are equal if and only if the variances of the two measurement error terms are equal. This is is a paper-and-pencil calculation. It is the basis of the next (and last) test you are asked to carry out.

(b) Carry out a Wald (not likelihood ratio) test of the null hypothesis that the variances of the two measurement error terms for Knowledge measurements 1 and 2 are equal. By the last calculation you did, this is equivalent to testing whether the two reliabilities are equal.

  i. What is the value of the chi-squared statistic? The answer is a number.
  ii. What are the degrees of freedom? The answer is a number.
  iii. What is the p-value? The answer is a number.
  iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
  v. What do you conclude about the reliabilties of the two measurements? Using $\alpha = 0.05$, do you have sufficient evidence to conclude that they are different? Answer Yes or No. If the answer is Yes, say which one seems to be more reliable. Of course the answer may not be Yes. If the answer is No, do *not* draw any conclusions about which measurement is more reliable.

Bring your log file and your results file to the quiz. You may be asked for numbers from your printouts, and you may be asked to hand them in. **There must be no error messages, and no notes or warnings about invalid data on your log file.**