STA 431 Assignment 10

Do this assignment in preparation for the quiz on Friday, March 27th. Bring your log and list file for Question 4. The other questions are practice for the quiz, and are not to be handed in.

1. Here is a regression model with one latent independent variable ξ, two dependent variables, and a single measurement of ξ (with error). Let

$$\begin{aligned} Y_1 &= \gamma_1 \xi + \zeta_1 \\ Y_2 &= \gamma_2 \xi + \zeta_2 \\ X &= \xi + \delta, \end{aligned}$$

   where δ, ξ, ζ_1 and ζ_2 are all independent normals with expected value zero, $Var(\xi) = \phi$, $Var(\zeta_1) = \psi_1$, $Var(\zeta_2) = \psi_2$, $Var(\delta) = \omega$, and the regression coefficients γ_1 and γ_2 are fixed constants. ξ is a latent variable.

   (a) Give the covariance matrix of the observed variables X, Y_1 and Y_2.

   (b) What are the parameters of this model? That is, give the parameter vector θ.

   (c) Is this model identified? Answer Yes or No and prove your answer.

2. Consider the following simple regression through the origin with measurement error in both the independent and dependent variables.

$$\begin{aligned} \eta &= \gamma \xi + \zeta \\ X_1 &= \xi + \delta_1, \\ X_2 &= \xi + \delta_2, \\ Y_1 &= \eta + \epsilon_1, \\ Y_2 &= \eta + \epsilon_2, \end{aligned}$$

   where ξ and η are latent variables, ζ, δ_1, δ_2, ϵ_1, ϵ_2 and ξ are normal random variables with expected value zero, $Var(\xi) = \phi$, $Var(\zeta) = \psi$, $Var(\delta_1) = \omega_{1,1}$, $Var(\epsilon_1) = \omega_{2,2}$, $Var(\delta_2) = \omega_{3,3}$, $Var(\epsilon_2) = \omega_{4,4}$, $Cov(\delta_1, \epsilon_1) = \omega_{1,2}$, and $Cov(\delta_2, \epsilon_2) = \omega_{3,4}$. If the covariance for a pair of random variables is not explicitly given, then they are independent. The regression coefficient γ is a fixed constant. The observed variables are X_1, X_2, Y_1 and Y_2.

   (a) What are the parameters of this model? That is, give the parameter vector θ.

   (b) Is this model identified? Answer Yes or No and prove your answer. Are calculations really necessary?

   (c) Is the model just identified, or is it over-identified? Choose one.

(d) How many over-identifying restrictions are there? The answer is a number.

(e) Calculate $\boldsymbol{\Sigma} = \sigma(\theta)$, and find the over-identifying restriction. What is it? Your answer is an equation in terms of $\sigma_{i,j}$ quantities (not model parameters).

3. Here is a multivariate multiple regression model with measurement error in the both the independent variables and the dependent variables. Let

$$\begin{aligned}
\boldsymbol{\eta} &= \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \\
\mathbf{X} &= \boldsymbol{\xi} + \boldsymbol{\delta} \\
\mathbf{Y} &= \boldsymbol{\eta} + \boldsymbol{\epsilon},
\end{aligned}$$

where $\boldsymbol{\eta}$ is an $m \times 1$ vector of latent dependent variables, $\boldsymbol{\xi}$ is a $p \times 1$ vector of latent independent variables with $V(\boldsymbol{\xi}) = \boldsymbol{\Phi}$, $\boldsymbol{\zeta}$ is an $m \times 1$ vector of error terms with $V(\boldsymbol{\zeta}) = \boldsymbol{\Psi}$, $\boldsymbol{\Gamma}$ is an $m \times p$ matrix of regression coefficients, $\mathbf{X}$ is a $p \times 1$ vector of observable variables, $\boldsymbol{\delta}$ is the measurement error in $\mathbf{X}$ with $V(\boldsymbol{\delta}) = \boldsymbol{\Omega}_\delta$, $\mathbf{Y}$ is an $m \times 1$ vector of observable variables, and $\boldsymbol{\epsilon}$ is the measurement error in $\mathbf{Y}$ with $V(\boldsymbol{\epsilon}) = \boldsymbol{\Omega}_\epsilon$. The exogenous variables $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ are all independent multivariate normals with expected value zero.

(a) Write the covariance matrix $\boldsymbol{\Sigma}$ of the observed variables as a partitioned matrix. What are the dimensions of $\boldsymbol{\Sigma}$?

(b) The parameter vector θ consists of all the unique elements of the matrices mentioned above. How many elements does θ have?

(c) Is this model identified in general, that is, without any further restrictions on the parameter matrices? Answer Yes or No and prove your answer. A simple numerical example is best for proving non-identification.

(d) In what way does this model *not* fit the double measurement design?

4. The file `pigs.data` comes from asking a sample of farmers the same questions twice (one month apart). X_1 and X_2 are reported number of breeding hogs on hand September 1, and Y_1 and Y_2 were reported number of pigs giving birth between June 1 and August 31. Please assume that by asking the questions a month apart (embedded in much bigger questionnaires), we have ensured that errors of measurement on the two interviews are independent. It is *not* safe to assume that answers to questions from the same interview have measurement errors that are independent.

(a) Do you see a connection with any other question from this assignment?

(b) Make a path diagram for these data.

(c) Would you expect the coefficient linking the two latent variables to be positive, or negative? Say which. Why?

(d) By making a path diagram, you have proposed a model (I believe that only one reasonable model is possible). Is the model identified? Answer Yes or No and establish the truth of your answer. You don't have to do any calculations for full marks.

(e) Use `proc calis` to estimate the parameters of your model by maximum likelihood. For each breeding pig on hand in September, what is the estimated number of pigs giving birth the following Spring/Summer? Your answer is a single number.

(f) What is the t (actually, Z) statistic associated with the last item? Your answer is a single number. What null hypothesis is being tested? Give your answer in the symbols of the LISREL model.

(g) What does the chisquare test indicate about the fit of the model? Choose one: The fit is

   i. Great
   ii. Not Great
   iii. Terrible

(h) How many over-identifying restrictions are there in this model? You can tell from the degrees of freedom.

(i) Can you tell exactly *what* the over-identifying restriction is? This is interesting because the test for goodness of fit is evaluating the model by testing whether the over-identifying restriction is correct.