STA 431s17 Assignment Ten[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 13) 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. The following model is centered, and has zero covariance between all pairs of exogenous variables including error terms.

$$\begin{aligned} Y_1 &= \gamma_1 X_1 + \gamma_2 X_2 + \epsilon_1 \\ Y_2 &= \beta Y_1 + \gamma_3 X_1 + \epsilon_2 \\ W_1 &= \lambda_1 X_1 + e_1 \\ W_2 &= \lambda_2 X_2 + e_2 \\ V_1 &= \lambda_3 Y_1 + e_3 \\ V_2 &= \lambda_4 Y_2 + e_4 \end{aligned}$$

   (a) Make a path diagram.

   (b) Referring to the general two-stage structural equation model on the formula sheet, write the model equations in matrix form. This means put symbols from the model above in the matrices. Also give the matrices $\boldsymbol{\Phi}_x$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$. The dimensions must be right for the specific model above.

2. Consider the general factor analysis model

$$\mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i,$$

   where $\boldsymbol{\Lambda}$ is a $k \times p$ matrix of factor loadings, the vector of factors $\mathbf{F}_i$ is a $p \times 1$ multivariate normal with expected value zero and covariance matrix $\boldsymbol{\Phi}$, and $\mathbf{e}_i$ is multivariate normal and independent of $\mathbf{F}_i$, with expected value zero and covariance matrix $\boldsymbol{\Omega}$. All covariance matrices are positive definite.

   (a) How do you know that $\mathbf{D}_i$ is multivariate normal?

   (b) Calculate the matrix of covariances between the observable variables $\mathbf{D}_i$ and the underlying factors $\mathbf{F}_i$.

   (c) Give the covariance matrix of $\mathbf{D}_i$. Show your work.

   (d) Because $\boldsymbol{\Phi}$ symmetric and positive definite, it has a square root matrix that is also symmetric. Using this, show that the parameters of the general factor analysis model are not identifiable.

(e) In an attempt to obtain a model whose parameters can be successfully estimated, let $\mathbf{\Omega}$ be diagonal (errors are uncorrelated) and set $\mathbf{\Phi}$ to the identity matrix (standardizing the factors). Show that the parameters of this revised model are still not identifiable. Hint: An orthogonal matrix $\mathbf{R}$ (corresponding to a rotation) is one satisfying $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$.

3. Let

$$\begin{aligned} D_1 &= \lambda_1 F_1 + e_1 \\ D_2 &= \lambda_2 F_2 + e_2 \\ D_3 &= \lambda_3 F_3 + e_3, \end{aligned}$$

where F_1, F_2, F_3, e_1, e_2 and e_3 are all independent with $F_j \sim N(0,1)$ and $e_j \sim N(0,\omega_j)$. All the expected values are zero. You can tell from the notation which variables are observable.

(a) Give the variance-covariance matrix of the observable variables.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

(c) Even though the parameters are not identifiable, the model itself is testable. That is, it implies a set of equality restrictions on the covariance matrix $\mathbf{\Sigma}$ that could be tested, and rejecting the null hypothesis would call the model into question. State the null hypothesis. Again, it is a statement about the $\sigma_{i,j}$ values.

4. Here is another factor analysis model. This one has a single underlying factor.

$$\begin{aligned} D_1 &= \lambda_1 F + e_1 \\ D_2 &= \lambda_2 F + e_2 \\ D_3 &= \lambda_3 F + e_3, \end{aligned}$$

where the factor and error terms are all independent, $F \sim N(0,1)$, $e_j \sim N(0,\omega_j)$, and λ_1, λ_2 and λ_3 are nonzero constants with $\lambda_1 > 0$.

(a) Give the variance-covariance matrix of the observed variables.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

5. Suppose we added another variable to the model of Question 4. That is, we add

$$D_4 = \lambda_4 F + e_4,$$

with assumptions similar to the ones of Question 4. Now suppose that $\lambda_2 = 0$, while the other factor loadings are non-zero.

(a) Is λ_2 identifiable? Justify your answer.

(b) Are the other factor loadings identifiable? Justify your answer.

6. Suppose we added a fifth variable to the model of Question 5. That is, we add

$$D_5 = \lambda_5 F + e_5,$$

with assumptions similar to the ones of Question 4. Now suppose that $\lambda_3 = \lambda_4 = 0$, while the other factor loadings are non-zero.

(a) Are λ_3 and λ_4 identifiable? Justify your answer.

(b) Are the other three factor loadings identifiable? Justify your answer.

(c) State the general pattern that is emerging here.

7. We now extend the model of Question 4 by adding a second factor. Let

$$\begin{aligned} D_1 &= \lambda_1 F_1 + e_1 \\ D_2 &= \lambda_2 F_1 + e_2 \\ D_3 &= \lambda_3 F_1 + e_3 \\ D_4 &= \lambda_4 F_2 + e_4 \\ D_5 &= \lambda_5 F_2 + e_5 \\ D_6 &= \lambda_6 F_2 + e_6, \end{aligned}$$

where all expected values are zero, $Var(e_i) = \omega_i$ for $i = 1, \ldots, 6$, $Var(F_1) = Var(F_2) = 1$, $Cov(F_1, F_2) = \phi_{12}$, the factors are independent of the error terms, and all the error terms are independent of each other. All the factor loadings are non-zero.

(a) Give the covariance matrix of the observable variables. Show the necessary work. A lot of the work has already been done.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

(c) Write the model in matrix form as $\mathbf{D} = \mathbf{\Lambda F} + \mathbf{e}$. That is give the matrices. For example, $\mathbf{D}$ is 6×1.

(d) Recall that a *rotation* matrix is any square matrix $\mathbf{R}$ satisfying $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$. Give a specific 2×2 rotation matrix $\mathbf{R}$ so that $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_2 = \mathbf{\Lambda R}$ yield the same $\mathbf{\Sigma} = cov(\mathbf{D})$. Hint: Use your answer to Question 7b.

(e) Suppose we add the conditions $\lambda_1 > 0$ and $\lambda_4 > 0$. Are the parameters identifiable now?

8. In Question 7, suppose we added just two variables along with the second factor. That is, we omit the equation for D_6, while keeping $\lambda_1 > 0$ and $\lambda_4 > 0$. Are the model parameters identifiable in this case? Answer Yes or No; show your work.

9. Let's add a third factor to the model of Question 7. That is, we keep the equation for D_6 and add

$$\begin{aligned} D_7 &= \lambda_7 F_3 + e_7 \\ D_8 &= \lambda_8 F_3 + e_8 \\ D_9 &= \lambda_9 F_3 + e_9 \end{aligned}$$

with $\lambda_1 > 0$, $\lambda_4 > 0$, $\lambda_7 > 0$ and other assumptions similar to the ones we have been using. Are the model parameters identifiable? You don't have to do any calculations if you see the pattern.

10. In this factor analysis model, the observed variables are *not* standardized, and the factor loading for D_1 is set equal to one. Let

$$\begin{aligned} D_1 &= F + e_1 \\ D_2 &= \lambda_2 F + e_2 \\ D_3 &= \lambda_3 F + e_3, \end{aligned}$$

where $F \sim N(0, \phi)$, e_1, e_2 and e_3 are normal and independent of F and each other with expected value zero, $Var(e_1) = \omega_1, Var(e_2) = \omega_2, Var(e_3) = \omega_3$, and λ_2 and λ_3 are nonzero constants.

(a) Calculate the variance-covariance matrix of the observed variables.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

11. We now extend the preceding model by adding another factor. Let

$$\begin{aligned} D_1 &= F_1 + e_1 \\ D_2 &= \lambda_2 F_1 + e_2 \\ D_3 &= \lambda_3 F_1 + e_3 \\ D_4 &= F_2 + e_4 \\ D_5 &= \lambda_5 F_2 + e_5 \\ D_6 &= \lambda_6 F_2 + e_6, \end{aligned}$$

where all expected values are zero, $Var(e_i) = \omega_i$ for $i = 1, \ldots, 6$,

$$cov \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix},$$

and $\lambda_2, \lambda_3, \lambda_5$ and λ_6 are nonzero constants.

(a) Give the covariance matrix of the observable variables. Show the necessary work. A lot of the work has already been done in Question 10.

(b) Are the model parameters identifiable? Answer Yes or No and prove your answer.

12. Let's add a third factor to the model of Question 11. That is, we add

$$\begin{aligned} D_7 &= F_3 + e_7 \\ D_8 &= \lambda_8 F_3 + e_8 \\ D_9 &= \lambda_9 F_3 + e_9 \end{aligned}$$

and

$$cov \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{12} & \phi_{22} & \phi_{23} \\ \phi_{13} & \phi_{23} & \phi_{33} \end{pmatrix},$$

with $\lambda_8 \neq 0$, $\lambda_9 \neq 0$ and so on. Are the model parameters identifiable? You don't have to do any calculations if you see the pattern.

13. The SAS part of this assignment is based on the Poverty Data. The data are given in the file poverty.data.txt . There is a link on the course web page in case the one in this document does not work. This data set contains information from a sample of 97 countries. In order, the variables include Live birth rate per 1,000 of population, Death rate per 1,000 of population, Infant deaths per 1,000 of population under 1 year old, Life expectancy at birth for males, Life expectancy at birth for females, and Gross National Product per capita in U.S. dollars. There is also a categorical variable representing location (continent), and finally the name of the country.

    This can be a very challenging and frustrating data set to work with, because correlated measurement errors produce negative variance estimates and other numerical problems almost everywhere you turn. To make your job easier, please confine your analyses to the following four variables:

    Life Expectancy: Average of life expectancy for males and life expectancy for females.

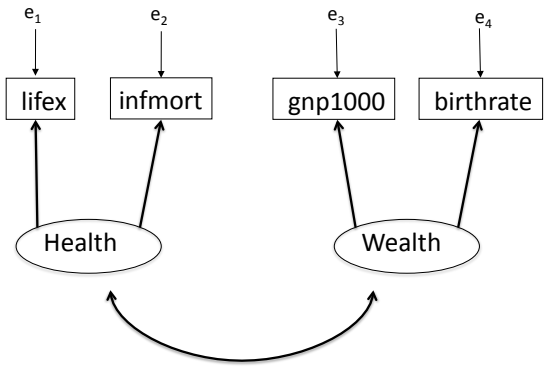
    Infant mortality rate.

    Birth rate.

    GNP/1000 = Gross national product in thousands of dollars. The re-scaling is a solution to numerical problems in fitting the model.

    You are not using all the variables in the data file, but you should read them all, because other ways of skipping variables are more trouble. The names of character-valued variables (the last two) must be followed by dollar signs ($).

    Here is a picture of a factor analysis model with 2 factors.



    The reason for making birth rate an indicator of wealth is that birth control costs money.

    (a) Fit the model with `proc calis`. My value of the Schwarz Bayesian Criterion (whatever that is) is 41.6961. Make sure to include the `pcorr` option, so you will get $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$. You will have to re-parameterize. Which of the two standard re-parameterizations should you choose? Suppose we are interested in the correlation between Health and Wealth.

    (b) What are the unknown parameters for this model? Give your answer in the form of a list of names from your SAS job.

(c) What rule tells you that the re-parameterization you have chosen results in a parameter vector that is identifiable (at least, it's identifiable in most of the parameter space)? Name the rule given in Lecture 20.

(d) Does this model fit the data adequately? Answer Yes or No, and back up your answer with two numbers from the printout: The value of a test statistic, and a p-value.

(e) Why does the goodness of fit test have one degree of freedom?

(f) What is the maximum likelihood estimate of the correlation between factors? The answer is a single number from the printout.

(g) Now fit a model with the other common re-parameterization, again including the `pcorr` option.

   i. Compare the two likelihood ratio tests for model fit. What do you see?
   ii. Compare the two $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ matrices. What do you see?
   iii. Give the maximum likelihood estimate of λ_2/λ_1 based on output from the *first* model. Can you find this number in the output from the second model?
   iv. Based on the output from the second model, give the maximum likelihood of the correlation between Health and Wealth. Can you find this number in the output from the first model?

(h) Finally, the high estimated correlation between factors from the first part of this question suggests that there might be just one underlying factor: wealth. Try a single-factor model and see if it fits. Locate the relevant chi-squared statistic, degrees of freedom and p-value. Do the estimated factor loadings make sense? What do you conclude? Do you like the one-factor model or the two-factor model?

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.** Warnings about missing data are okay. The USSR did not co-operate.