STA 2101f19 Assignment Seven[1]

This assignment is about the double measurement design. See lecture Slide Sets 18, 19 and 20, and Section 0.11 (pages 63-104) in Chapter Zero. The non-computer questions on this assignment are for practice, and will not be handed in. For the R part of this assignment (Question 4) please bring hard copy of your input and output to the quiz.

1. The point of this question is that when the parameters of a model are identifiable, the number of covariance structure equations minus the number of parameters equals the number of model-induced equality constraints on $\boldsymbol{\Sigma}$. It is these equality constraints that are being tested by the chi-squared test for goodness of fit.

   In the lecture notes, look at the matrix formulation and discussion of double measurement regression starting on Slide 6 of lecture unit 19. The latent vector $\mathbf{X}_i$ is $p \times 1$, and the latent vector $\mathbf{Y}_i$ is $q \times 1$. As usual, expected values and intercepts are not identifiable, so confine your attention to $\boldsymbol{\Sigma} = [\sigma_{ij}]$, the covariance matrix of the observable data.

   (a) Here's something that will help with the calculations in this problem. If a covariance matrix is $n \times n$,

      i. How many unique covariances are there?

      ii. How many unique variances and covariances total are there? Factor and simplify.

   (b) For the present problem, what are the dimensions of $\boldsymbol{\Sigma}$? Give the number of rows and the number of columns. It's an expression in p and q.

   (c) How many unique variances and covariances (σ_{ij} quantities) are there in $\boldsymbol{\Sigma}$ when there are no model-induced constraints? The answer is an expression in p and q.

   (d) Denoting $cov(\mathbf{F}_i)$ by $\boldsymbol{\Phi} = [\phi_{ij}]$, how many unique variances and covariances (ϕ_{ij} quantities) are there in $\boldsymbol{\Phi} = cov(\mathbf{F}_i)$ if there are no model-induced equality constraints? The answer is an expression in p and q.

   (e) In total, how many unknown parameters are there in the Stage One parameter matrices $\boldsymbol{\Phi}_x$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\Psi}$? The answer is an expression in p and q. Is this the same as your last answer? If so, it means that at the first stage, if the parameters are identifiable from $\boldsymbol{\Phi}$, they are *just identifiable* from $\boldsymbol{\Phi}$.

   (f) Still in Stage One (the latent variable model), show the details of how the parameter matrices $\boldsymbol{\Phi}_x$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\Psi}$ can be recovered from $\boldsymbol{\Phi}$. Start by calculating $\boldsymbol{\Phi}$ as a function of $\boldsymbol{\Phi}_x$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\Psi}$. You have shown that the function relating $\boldsymbol{\Phi}$ to $(\boldsymbol{\Phi}_x, \boldsymbol{\beta}_1, \boldsymbol{\Psi})$ is one-to-one (injective).

(g) In Stage Two (the measurement model), the parameters are in the matrices $\mathbf{\Phi}$, $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$. How many unique parameters are there? The answer is an expression in p and q.

(h) By inspecting the expression for $\mathbf{\Sigma}$ on Slide 11 of lecture unit 19, state the number of equality constraints that are imposed on $\mathbf{\Sigma}$ by the model. The answer is an expression in p and q.

(i) Show that the number of parameters plus the number of constraints is equal to the number of unique variances and covariances in $\mathbf{\Sigma}$. This is a brief calculation using your answers to 1c and the last two questions.

2. Here is a one-stage formulation of the double measurement regression model. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned}
\mathbf{W}_{i,1} &= \mathbf{X}_i + \mathbf{e}_{i,1} \\
\mathbf{V}_{i,1} &= \mathbf{Y}_i + \mathbf{e}_{i,2} \\
\mathbf{W}_{i,2} &= \mathbf{X}_i + \mathbf{e}_{i,3}, \\
\mathbf{V}_{i,2} &= \mathbf{Y}_i + \mathbf{e}_{i,4}, \\
\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i
\end{aligned}$$

where

$\mathbf{Y}_i$ is a $q \times 1$ random vector of latent response variables. Because q can be greater than one, the regression is multivariate.

$\boldsymbol{\beta}$ is an $q \times p$ matrix of unknown constants. These are the regression coefficients, with one row for each response variable and one column for each explanatory variable.

$\mathbf{X}_i$ is a $p \times 1$ random vector of latent explanatory variables, with expected value zero and variance-covariance matrix $\mathbf{\Phi}_x$, a $p \times p$ symmetric and positive definite matrix of unknown constants.

$\boldsymbol{\epsilon}_i$ is the error term of the latent regression. It is a $q \times 1$ random vector with expected value zero and variance-covariance matrix $\mathbf{\Psi}$, a $q \times q$ symmetric and positive definite matrix of unknown constants.

$\mathbf{W}_{i,1}$ and $\mathbf{W}_{i,2}$ are $p \times 1$ observable random vectors, each representing $\mathbf{X}_i$ plus random error.

$\mathbf{V}_{i,1}$ and $\mathbf{V}_{i,2}$ are $q \times 1$ observable random vectors, each representing $\mathbf{Y}_i$ plus random error.

$\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,4}$ are the measurement errors in $\mathbf{W}_{i,1}, \mathbf{V}_{i,1}, \mathbf{W}_{i,2}$ and $\mathbf{V}_{i,2}$ respectively. Joining the vectors of measurement errors into a single long vector $\mathbf{e}_i$, its covari-

ance matrix may be written as a partitioned matrix

$$cov(\mathbf{e}_i) = cov\begin{pmatrix} \mathbf{e}_{i,1} \\ \mathbf{e}_{i,2} \\ \mathbf{e}_{i,3} \\ \mathbf{e}_{i,4} \end{pmatrix} = \left(\begin{array}{c|c|c|c} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} & \mathbf{0} & \mathbf{0} \\ \hline \boldsymbol{\Omega}_{12}^\top & \boldsymbol{\Omega}_{22} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega}_{33} & \boldsymbol{\Omega}_{34} \\ \hline \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega}_{34}^\top & \boldsymbol{\Omega}_{44} \end{array}\right) = \boldsymbol{\Omega}.$$

In addition, the matrices of covariances between $\mathbf{X}_i, \boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are all zero.

Collecting $\mathbf{W}_{i,1}$, $\mathbf{W}_{i,2}$, $\mathbf{V}_{i,1}$ and $\mathbf{V}_{i,2}$ into a single long data vector $\mathbf{D}_i$, we write its variance-covariance matrix as a partitioned matrix:

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c|c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \hline & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \hline & & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \hline & & & \boldsymbol{\Sigma}_{44} \end{array}\right),$$

where the covariance matrix of $\mathbf{W}_{i,1}$ is $\boldsymbol{\Sigma}_{11}$, the covariance matrix of $\mathbf{V}_{i,1}$ is $\boldsymbol{\Sigma}_{22}$, the matrix of covariances between $\mathbf{W}_{i,1}$ and $\mathbf{V}_{i,1}$ is $\boldsymbol{\Sigma}_{12}$, and so on.

(a) Write the elements of the partitioned matrix $\boldsymbol{\Sigma}$ in terms of the parameter matrices of the model. Be able to show your work for each one.

(b) Prove that all the model parameters are identifiable by solving the covariance structure equations.

(c) Give a Method of Moments estimator of $\boldsymbol{\Phi}_x$. Remember, your estimator cannot be a function of any unknown parameters. For a particular sample, will your estimate be in the parameter space? Mine is.

(d) Give a Method of Moments estimator for $\boldsymbol{\beta}$. Remember, your estimator cannot be a function of any unknown parameters. There is more than one correct answer. How do you know your estimator is consistent?

3. Question 4 (the R part of this assignment) will use the *Pig Birth Data*. As part of a much larger study, farmers filled out questionnaires about various aspects of their farms. Some questions were asked twice, on two different questionnaires several months apart. Buried in all the questions were

   - Number of breeding sows (female pigs) at the farm on June 1st
   - Number of sows giving birth later that summer.

   There are two readings of these variables, one from each questionnaire. We will assume (maybe incorrectly) that because the questions were buried in a lot of other material and were asked months apart, that errors of measurement are independent between the two questionnaires. However, errors of measurement might be correlated within a questionnaire.

(a) Propose a reasonable model for these data, using the usual notation. Give all the details. You may assume normality if you wish. Remember, measurement error could be correlated within questionnaires.

(b) Make a path diagram of the model you have proposed.

(c) Of course it is hopeless to identify the expected values and intercepts, so we will concentrate on the covariance matrix. Calculate the covariance matrix of one observable data vector $\mathbf{D}_i$. Compare your answer to 2a.

(d) Even though you have a general result that applies to this case, prove that all the parameters in the covariance matrix are identifiable.

(e) If there are any equality constraints on the covariance matrix, say what they are.

(f) Based on your answer to the last question, how many degrees of freedom should there be in the chi-squared test for model fit? Does this agree with your answer to Question 1h?

(g) Give a consistent estimator of β that is *not* the MLE, and explain why it's consistent. You may use the consistency of sample variances and covariances without proof. Your estimator *must not* be a function of any unknown parameters.

4. The Pig Birth Data are given in the file `openpigs.data.txt`. No doubt you will have to edit the data file to strip off the information at the top. There are $n = 114$ farms; please verify that you are reading the correct number of cases.

(a) Start by reading the data and producing a correlation matrix of all the observable variables.

(b) Use `lavaan` to fit your model, and look at `summary`. If you experience numerical problems you are doing something differently from the way I did it. When I fit a good model everything was fine. When I fit a poor model there was trouble. Just to ensure we are fitting the same model, my log likelihood (obtained with the `logLik` function) was `-1901.717`.

(c) Does your model fit the data adequately? Answer Yes or No and give three numbers: a chi-squared statistic, the degrees of freedom, and a p-value.

(d) For each additional breeding sow present in September, estimated number giving birth that summer goes up by _____. Your answer is a single number from `summary`. It is not an integer.

(e) Using your answer to Question 3g, give a *numerical* version of your consistent estimate of β. How does it compare to the MLE?

(f) Give a large-sample confidence interval for your answer to 4d. Note that $\sqrt{n}$ is already built into the inverse of the Hessian, so you don't need multiply by it again. Using all the accuracy available, my lower confidence limit is `0.6510766`.

(g) Recall that reliability of a measurement is the proportion of its variance that does *not* come from measurement error. What is the estimated reliability of the number of breeding sows from questionnaire two? The answer is a number, which you could get with a calculator and the output of `summary`.

(h) Is there evidence of correlated measurement error within questionnaires? Answer Yes or No and give some numbers from the results file to support your conclusion.

(i) The answer to that last question was based on two separate tests. Though it is already pretty convincing, conduct a *single* Wald (not likelihood ratio) test of the two null hypotheses simultaneously. Give the Wald chi-squared statistic, the degrees of freedom and the p-value. What do you conclude? Is there evidence of correlated measurement error, or not?

(j) The double measurement design allows the measurement error covariance matrices $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ to be unequal. Carry out a Wald test to see whether the two covariance matrices are equal or not.

   i. Give the Wald chi-squared statistic, the degrees of freedom and the p-value. What do you conclude? Is there evidence that the two measurement error covariance matrices are unequal?

   ii. There is evidence that one of the measurements is less accurate on one questionnaire than the other. Which one is it? Give the Wald chi-squared statistic, the degrees of freedom and the p-value. I did two tests here; only one of them was significant.

Bring a printout with your R input and output to the quiz. Please remember that while the questions may appear in comment statements, answers and interpretation may not, except for numerical answers generated by R.