STA 431s15 Assignment Six[1]

The non-computer questions on this assignment are for practice, and will not be handed in. For the SAS part of this assignment (Question 4) please bring your log file and your output file to the quiz. There may be one or more questions about them, and you may be asked to hand the printouts in with the quiz.

1. In the lecture notes, look at the matrix formulation of double measurement regression. As usual, expected values and intercepts are not identifiable, so confine your attention to the covariance matrix.

   (a) How many unknown parameters appear in $\boldsymbol{\Sigma}$? The answer is an expression in p and q.

   (b) How many unique variances and covariances are there in $\boldsymbol{\Phi} = V(\mathbf{F}_i)$? The answer is an expression in p and q.

   (c) In total, how many unknown parameters are there in the Stage One 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 $\boldsymbol{\Phi}$.

   (d) 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 $\boldsymbol{\Phi}$.

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

   (f) How many unique variances and covariances are there in $\boldsymbol{\Sigma}$? The answer is an expression in p and q.

   (g) How many equality constraints are imposed on $\boldsymbol{\Sigma}$ by the model? The answer is an expression in p and q.

   (h) Show that the number of parameters plus the number of constraints is equal to the number of unique variances and covariances in $\boldsymbol{\Sigma}$. This is a brief calculation using your earlier answers.

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

[1] This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/431s15

$\mathbf{Y}_i$ 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 $\boldsymbol{\Phi}$, 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 $\boldsymbol{\Psi}$, 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.

$\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,1}$ 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 covariance matrix may be written as a partitioned matrix

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

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 $\boldsymbol{\Phi}$. There is more than one reasonable answer. Remember, your estimator cannot be a function of any unknown parameters, or you get a zero. For a particular sample, will your estimate be in the parameter space? Mine is.

(d) Give a Method of Moments estimator of $\boldsymbol{\beta}$. Remember, your estimator cannot be a function of any unknown parameters, or you get a zero. How do you know your estimator is consistent? You may use $\widehat{\boldsymbol{\Sigma}} \overset{a.s.}{\rightarrow} \boldsymbol{\Sigma}$ without proof.

3. Question 4 (the SAS 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) Write down a reasonable model for these data, using the usual notation. Give all the details. You may assume normality if you wish.

   (b) 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 $\mathbf{D}_i$.

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

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

   (e) Based on your answer to the last question, how many degrees of freedom should there be in the chisquare test for model fit? Does this agree with your answer to Question 1g?

   (f) 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, or you get a zero on this part.

4. The Pig Birth Data are given in the file `openpigs.data.txt`. There is a link on the course web page in case the one in this document does not work. Note there are $n = 114$ farms, so please verify that you are reading the correct number of cases. Use the `firstobs` option in your `infile` statement.

   (a) Start by reading the data and then running `proc corr` to produce a correlation matrix (with tests) of all the observable variables.

   (b) Use `proc calis` to fit your model. Please use the `pshort nostand vardef=n pcorr` options. 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.

   (c) 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.

   (d) For each breeding sow present in September, what is the predicted number giving birth that summer? Your answer is a single number from the list file. It is not an integer.

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

   (f) 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 number of breeding sows from questionnaire two? The answer is a number, which you get with a calculator and the output file.

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

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