STA 431s23 Assignment Seven[1]

For the Quiz on Friday March 10th, please bring a printout of your full R input and output for Question 3. The other problems are not to be handed in. They are practice for the Quiz.

---

1. 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 a $q \times p$ 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 variance-covariance matrix $\boldsymbol{\Phi}_x$.

   $\boldsymbol{\epsilon}_i$ is the error term of the latent regression. It is a $q \times 1$ random vector with variance-covariance matrix $\boldsymbol{\Psi}$.

   $\mathbf{w}_{i,1}$ 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 covariance 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.

---

[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/431s23

Collecting $\mathbf{w}_{i,1}$, $\mathbf{v}_{i,1}$, $\mathbf{w}_{i,2}$ 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. Once you have solved for a parameter matrix, ou may use it in later solutions.

(c) Give a Method of Moments estimator of $\boldsymbol{\Phi}_x$. 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 for $\boldsymbol{\beta}$. 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 $\widehat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}$ without proof.

2. For the double measurement regression model of Question 1,

(a) How many unknown parameters appear in the covariance matrix of the observable variables?

(b) How many unique variances and covariances are there in the covariance matrix of the observable variables? This is also the number of covariance structure equations.

(c) How many equality constraints does the model impose on the covariance matrix of the observable variables? What are they?

(d) Does the number of covariance structure equations minus the number of parameters equal the number of constraints?

3. 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.

The Pig Birth Data are given in the file `openpigs2.data.txt`.

(a) Start by reading the data. There are $n = 114$ farms; please verify that you are reading the correct number of cases.

(b) Use the `var` function to produce a sample covariance matrix of all the observable variables. Don't worry about n versus $n - 1$.

(c) Make a path diagram of the double measurement model for these data.

(d) Give the details of your model in centered form, supplying any necessary notation.

(e) Use `lavaan` to fit your model. 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 verify that we are fitting the same model, my estimate of the variance of the latent exogenous variable is 357.145.

(f) 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. Do the degrees of freedom agree with your answer to Question 2?

(g) If the number of breeding sows present in September increases by one, what happens to the estimated number giving birth that summer? You answer is based on a single number from the output of `summary`. It is not an integer.

(h) Using your answer to Question 1d and the output of `var`, give a method of moments estimate of β. How does it compare to the MLE?

(i) Give a large-sample confidence interval for your answer to 3g. I used `parameterEstimates` to do it the easy way.

(j) 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? There are two numbers, one for questionnaire one and another for questionnaire two. You could do this with a calculator and the output of `summary`, but I did it with `:=` in the model string.

(k) It would be inconvenient at best to get confidence intervals for reliability with a calculator. Obtain confidence intervals for the reliabilities in Question 3j. Check `parameterEstimates`. For the record, the standard errors are calculated using the delta method, described in Appendix A.

(l) Is there evidence of correlated measurement error within questionnaires? Answer Yes or No and give some numbers from the output of `summary` to support your conclusion.

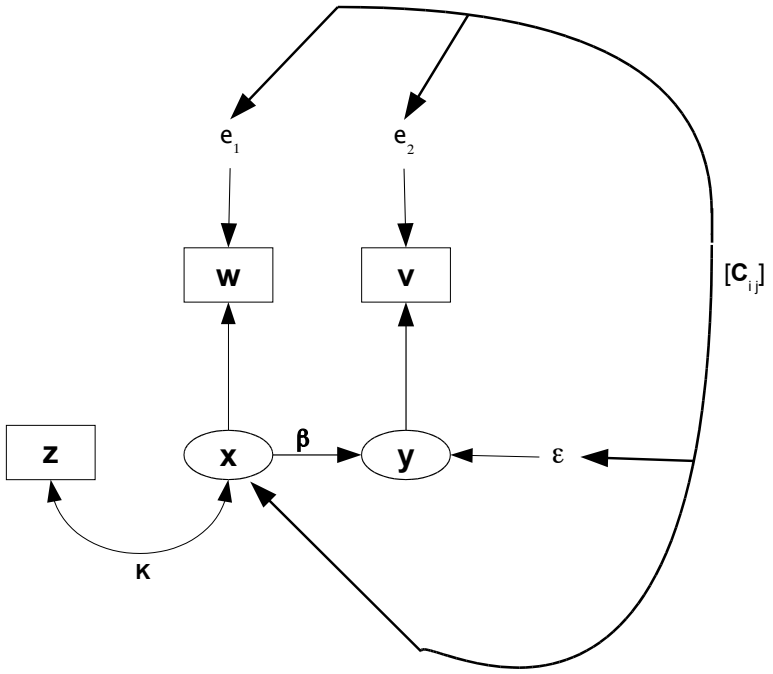
(m) In the double measurement design, two measurements of a latent variable are equally precise if their error variances are the same. We want to know whether the two measurements of number of breeding sows are equally precise, and also whether the two measurements of the number giving birth are equally precise.

   i. Give the null hypothesis for testing both comparisons at the same time.

   ii. Carry out the Wald test. This is a question about variances, so you can't trust the normal theory $\widehat{\mathbf{V}}_n$. Bootstrap it.

   iii. If the overall test was statistically significant, follow it up with separate tests of the two comparisons. Be able to report the test statistic, degrees of freedom, and p-value. Draw directional conclusions if appropriate, being guided by $\alpha = 0.05$, but never mentioning it.

4. Double measurement is not the only solution to measurement error in regression. Instrumental variables can take care of measurement error and omitted variables at the same time. The following path diagram not only betrays my lack of control over my drawing program, it shows a matrix version of instrumental variables with single measurement of both $\mathbf{x}$ and $\mathbf{y}$ and also massive covariance connecting $\mathbf{x}$ with all the error terms, and all the error terms with one another. These covariances arise from omitted variables.



In centered form (meaning without expected values or intercepts) the model equations are (independently for $i = 1, \ldots, n$),

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\epsilon}_i \\ \mathbf{w}_i &= \mathbf{x}_i + \mathbf{e}_{i,1} \\ \mathbf{v}_i &= \mathbf{y}_i + \mathbf{e}_{i,2}. \end{aligned}$$

As usual, $\mathbf{x}_i$ is $p \times 1$ and $\mathbf{y}_i$ is $q \times 1$. The $p \times 1$ vector of instrumental variables $\mathbf{z}_i$ has covariance matrix $\boldsymbol{\Phi}_z$. The covariances among the exogenous variables and error terms are given in the partitioned symmetric matrix

$$\mathbf{C} = cov\begin{pmatrix} \mathbf{x}_i \\ \hline \boldsymbol{\epsilon}_i \\ \hline \mathbf{e}_{i,1} \\ \hline \mathbf{e}_{i,2} \end{pmatrix} = \left(\begin{array}{c|c|c|c} \boldsymbol{\Phi}_x & \mathbf{C}_{12} & \mathbf{C}_{13} & \mathbf{C}_{14} \\ \hline & \boldsymbol{\Psi} & \mathbf{C}_{23} & \mathbf{C}_{24} \\ \hline & & \boldsymbol{\Omega}_1 & \mathbf{C}_{34} \\ \hline & & & \boldsymbol{\Omega}_2 \end{array}\right).$$

Finally, the $p \times p$ matrix $\mathbf{K} = cov(\mathbf{z}_i, \mathbf{x}_i)$ has an inverse.

(a) What are the dimensions of $\mathbf{C}_{23}$ (number of rows and columns)?

(b) Now we will count unknown parameters. It will help to recall that the number of unique elements in a $k \times k$ symmetric matrix is $k(k+1)/2$.

i. How many parameters are in the matrix $\mathbf{C}$?

ii. How many parameters are in the matrix $\mathbf{\Phi}_z$?

iii. How many parameters are in the matrix $\mathbf{K}$?

(c) Letting

$$\mathbf{\Sigma} = cov\left(\begin{array}{c}\mathbf{z}_i \\ \hline \mathbf{w}_i \\ \hline \mathbf{v}_i\end{array}\right) = \left(\begin{array}{c|c|c}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{13} \\ \hline & \mathbf{\Sigma}_{22} & \mathbf{\Sigma}_{23} \\ \hline & & \mathbf{\Sigma}_{33}\end{array}\right),$$

how many covariance structure equations are there?

(d) Is there any way this model passes the test of the parameter count rule?

(e) The naive model is $\mathbf{v}_i = \boldsymbol{\beta}\mathbf{w}_i + \boldsymbol{\epsilon}_i$, with zero covariance between $\mathbf{w}_i$ and $\boldsymbol{\epsilon}_i$. Ignoring $\mathbf{z}_i$, which is what people would do in practice, give a method of moments estimator of $\boldsymbol{\beta}$ (which is also MLE and least squares), in terms of $\widehat{\mathbf{\Sigma}}_{ij}$ matrices. Call it $\widehat{\boldsymbol{\beta}}_{bad}$.

(f) To what target does $\widehat{\boldsymbol{\beta}}_{bad}$ converge in probability under the true model? My answer is

$$\widehat{\boldsymbol{\beta}}_{bad} \stackrel{p}{\to} \left(\boldsymbol{\beta}\mathbf{\Phi}_x + \boldsymbol{\beta}\mathbf{C}_{13} + \mathbf{C}_{21} + \mathbf{C}_{23} + \mathbf{C}_{41} + \mathbf{C}_{43}\right)\left(\mathbf{\Phi}_x + \mathbf{C}_{13} + \mathbf{C}_{31} + \mathbf{\Omega}_1\right)^{-1}.$$

(g) We can do better than this. Propose another estimator of $\boldsymbol{\beta}$, and show that it converges in probability to $\boldsymbol{\beta}$ in most of the parameter space. Where does it fail?

(h) Why have you also just shown that $\boldsymbol{\beta}$ is identifiable except for a set of volume zero in the parameter space?

5. When instrumental variables are not available, sometimes identifiability can be obtained by adding more response variables to the model. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} W_i &= X_i + e_i \\ Y_{i,1} &= \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \beta_2 X_i + \epsilon_{i,2} \end{aligned}$$

where X_i, e_i, $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are all independent, $Var(X_i) = \phi > 0$, $Var(e_i) = \omega > 0$, $Var(\epsilon_{i,1}) = \psi_1 > 0$, $Var(\epsilon_{i,2}) = \psi_2 > 0$, and all the expected values are zero. The explanatory variable X_i is latent, while W_i, $Y_{i,1}$ and $Y_{i,2}$ are observable.

(a) Make a path diagram for this model

(b) What are the unknown parameters in this model?

(c) Let $\boldsymbol{\theta}$ denote the vector of Greek-letter unknowns that apear in the covariance matrix of the observable data. What is $\boldsymbol{\theta}$?

(d) Does this model pass the test of the Parameter Count Rule? Answer Yes or No and give the numbers.

(e) Calculate the variance-covariance matrix of the observable variables. Show your work.

(f) The parameter of primary interest is β_1. Is β_1 identifiable at points in the parameter space where $\beta_1 = 0$? Why or why not?

(g) Is ω identifiable where $\beta_1 = 0$?

(h) Give a simple numerical example to show that β_1 is not identifiable at points in the parameter space where $\beta_1 \neq 0$ and $\beta_2 = 0$.

(i) Is β_1 identifiable at points in the parameter space where $\beta_2 \neq 0$? Answer Yes or No and prove your answer.

(j) Show that the entire parameter vector is identifiable at points in the parameter space where $\beta_1 \neq 0$ and $\beta_2 \neq 0$.

(k) Propose a Method of Moments estimator of β_1.

(l) How do you know that your estimator cannot be consistent in a technical sense?

(m) For what points in the parameter space will your estimator converge in probability to β_1?

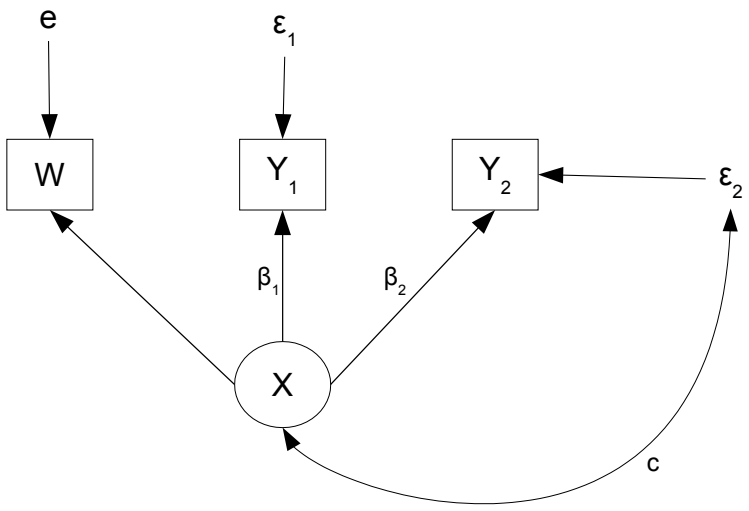
(n) How do you know that your Method of Moments estimator is also the Maximum Likelihood estimator (assuming normality)?

(o) Explain why the likelihood rato test of $H_0 : \beta_1 = 0$ will fail. Hint: What will happen when you try to locate $\widehat{\theta}_0$?

(p) Since the parameter of primary interest is β_1, it's important to be able to test $H_0 : \beta_1 = 0$. So at points in the parameter space where $\beta_2 \neq 0$, what *two* equality constraints on the elements of $\boldsymbol{\Sigma}$ are implied by $H_0 : \beta_1 = 0$? Why is this unexpected?

(q) Assuming $\beta_1 \neq 0$ and $\beta_2 \neq 0$, you can use the model to deduce more than one testable *inequality* constraint on the variances and covariances. Give at least one example.

6. In the model of Question 5, suppose that X and the extra response variable Y_2 are influenced by common omitted variables, so that there is non-zero covariance between X and ϵ_2. Here is a path diagram.



(a) How do you know that the full set of parameters (that is, the ones that appear in $\mathbf{\Sigma}$) cannot possibly be identifiable in most of the parameter space?

(b) Calculate the variance-covariance matrix of the observable variables. How does your answer compare to the one in Question 5?

(c) Primary interest is still in β_1. Propose a Method of Moments estimator of β_1. Is it the same as the one in Question 5, or different?

(d) For what set of points in the current parameter space is β_1 identifiable?

(e) If you had data in hand, what null hypothesis could you test about the σ_{ij} quantities to verify the identifiability of β_1? My null hypothesis has two equal signs.

(f) Suppose you want to test $H_0 : \beta_1 = 0$, which is likely the main objective.

i. If you rejected the null hypothesis in Question 6e, what null hypothesis would you test about the σ_{ij} quantities to test $H_0 : \beta_1 = 0$? My null hypothesis has two equal signs.

ii. If you failed to reject the null hypothesis in Question 6e, could you still test $H_0 : \beta_1 = 0$? What is the test on σ_{ij} quantities in this case?

iii. If you rejected $H_0 : \beta_1 = 0$, naturally you would want to state whether β_1 is positive or negative. Is this possible?

Please bring a printout of your full R input and output for Question 3 to the quiz.