

STA 431s17 Assignment Seven¹

This assignment is about the double measurement design. See lecture Slide Sets 11 and 12, and Section 0.11 (pages 51-61) in Chapter Zero. 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 hard copy of your log file and your results 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. 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 Σ . 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 25. 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 $\Sigma = [\sigma_{ij}]$, 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? Factor and simplify.
 - ii. How many unique variances and covariances total are there? Factor and simplify.
- (b) What are the dimensions of Σ ? 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 Σ when there are no model-induced constraints? The answer is an expression in p and q .
- (d) List the parameter matrices that appear in Σ .
- (e) Denoting $cov(\mathbf{F}_i)$ by $\Phi = [\phi_{ij}]$, how many unique variances and covariances (ϕ_{ij} quantities) are there in $\Phi = cov(\mathbf{F}_i)$ if there are no model-induced equality constraints? The answer is an expression in p and q .
- (f) In total, how many unknown parameters are there in the Stage One matrices Φ_x , β_1 and Ψ ? 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 Φ , they are *just identifiable* from Φ .

¹This 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](https://creativecommons.org/licenses/by-sa/3.0/). Use any part of it as you like and share the result freely. The L^AT_EX source code is available from the course website: <http://www.utstat.toronto.edu/~brunner/oldclass/431s17>

- (g) Still in Stage One (the latent variable model), show the details of how the parameter matrices Φ_x , β_1 and Ψ can be recovered from Φ . Start by calculating Φ as a function of Φ_x , β_1 and Ψ . You have shown that the function relating Φ to (Φ_x, β_1, Ψ) is one-to-one.
- (h) In Stage Two (the measurement model), the parameters are in the matrices Φ , Ω_1 and Ω_2 . How many unique parameters are there? The answer is an expression in p and q .
- (i) By inspecting the expression for Σ on Slide 30, state the number of equality constraints that are imposed on Σ by the model. The answer is an expression in p and q .
- (j) Show that the number of parameters plus the number of constraints is equal to the number of unique variances and covariances in Σ . 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 $i = 1, \dots, n$, let

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

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.

β 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 Φ_x , a $p \times p$ symmetric and positive definite matrix of unknown constants.

ϵ_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 Ψ , 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.

$\mathbf{e}_{i,1}, \dots, \mathbf{e}_{i,4}$ 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

$$\text{cov}(\mathbf{e}_i) = \text{cov} \begin{pmatrix} \mathbf{e}_{i,1} \\ \mathbf{e}_{i,2} \\ \mathbf{e}_{i,3} \\ \mathbf{e}_{i,4} \end{pmatrix} = \begin{pmatrix} \Omega_{11} & \Omega_{12} & \mathbf{0} & \mathbf{0} \\ \Omega_{12}^\top & \Omega_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Omega_{33} & \Omega_{34} \\ \mathbf{0} & \mathbf{0} & \Omega_{34}^\top & \Omega_{44} \end{pmatrix} = \Omega.$$

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:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ & & \Sigma_{33} & \Sigma_{34} \\ & & & \Sigma_{44} \end{pmatrix},$$

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

- (a) Write the elements of the partitioned matrix Σ 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 $\boldsymbol{\Phi}_x$. 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? Use $\hat{\Sigma} \xrightarrow{p} \Sigma$.
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) Propose a reasonable model for these data, using the usual notation. Give all the details. You may assume normality if you wish.
 - (b) Make a path diagram of the model you have proposed.
 - (c) Write the model equations again, this time in centered form. The little c symbols above the variables can be invisible.
 - (d) 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 .
 - (e) Even though you have a general result that applies to this case, prove that all the parameters in the covariance matrix are identifiable.
 - (f) If there are any equality constraints on the covariance matrix, say what they are.
 - (g) 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 1i?
 - (h) 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](#). Use the `firstobs` option in your `infile` statement to skip the first few lines. This is preferable to stripping the data file of documentation. There are $n = 114$ farms; please verify that you are reading the correct number of cases.
- (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. Just to verify that we are fitting the same model, my value of the Akaike Information Criterion (which we're not using) is 18.0871.
 - (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 breeding sow present in September, what is the predicted number giving birth that summer? Your answer is a single number from the results file. It is not an integer.
- (e) Using your answer to Question 3h, the results file and a calculator, give a *numerical* version of your consistent estimate of β . How does it compare to the MLE?
- (f) Since maximum likelihood estimates are asymptotically normal (approximately normal for large samples), a large-sample confidence interval is $\hat{\theta} \pm 1.96se$, where se is the standard error (estimated standard deviation) of $\hat{\theta}$. Give a large-sample confidence interval for your answer to 4d.
- (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 number of breeding sows from questionnaire two? The answer is a number, which you get with a calculator and the output file.
- (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. The SAS program `bmi3.sas` has an example of how to do a Wald test.
 - i. 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?
 - ii. Find two examples of $Z^2 \sim \chi^2(1)$ from the output for this question. Locate the tests and verify that the one-sided p -value from the χ^2 test equals the two-sided p -value from the Z test.
- (j) The double measurement design allows the measurement error covariance matrices Ω_1 and Ω_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.

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 are lots of **There must be no error messages, and no notes or warnings about invalid data on your log file.**