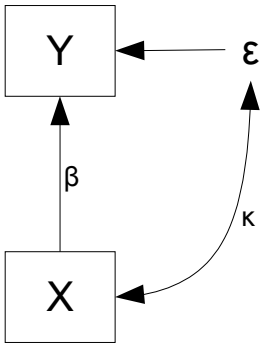
STA 431s15 Assignment Seven[1]

The non-computer questions on this assignment are practice for Term Test 2 and the final exam; they will not be handed in. There will be no SAS on Term Test 2. The SAS part of this assignment (Questions 7 and 8) are for Quiz Seven on Friday March 20. Please bring your log files and your output files to the quiz. There will be one or more questions about them, and you will be asked to hand printouts in with the quiz.

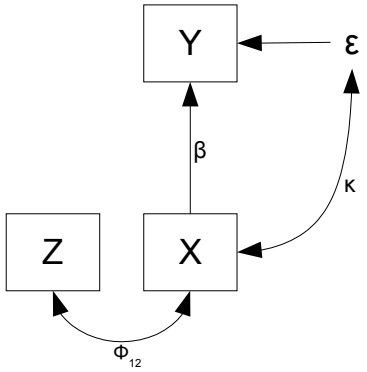
1. Since the error term ϵ_i in a regression equation represents "everything else," omission of explanatory variables that are correlated with the explanatory variables in the model will induce a non-zero covariance between the error term and the explanatory variables in the model. Throughout this question, we will set expected values and intercepts aside, and focus on the covariance matrix. Thus, parameters will be called "identifiable" if and only if they are identifiable from the covariance matrix of the observable data.

   (a) Here is an example for simple regression with no measurement error.



   i. Give the model equation in centered form. The centering can be invisible.
   ii. What is the parameter vector $\boldsymbol{\theta}$ for this model?
   iii. Does this model pass the test of the Parameter Count Rule? Answer Yes or No and give the numbers.
   iv. Calculate the covariance matrix $\boldsymbol{\Sigma}$ of the observable data vector.
   v. Is the parameter β identifiable? Answer Yes or No. If the answer is Yes, prove it. If the answer is No, give a simple numerical example of two parameter vectors with *different* β values, yielding the same covariance matrix $\boldsymbol{\Sigma}$ of the observable data.

(b) Now we add an instrumental variable Z.



i. Give the model equation in centered form. The centering can be invisible.

ii. What is the parameter vector $\boldsymbol{\theta}$ for this model?

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

iv. Calculate the covariance matrix $\boldsymbol{\Sigma}$ of the observable data vector.

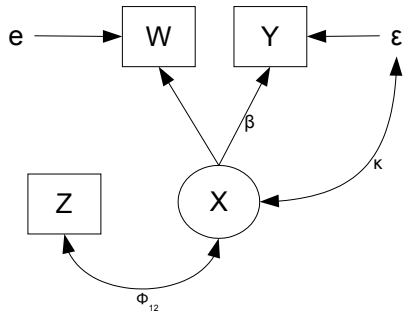
v. Is the entire parameter $\boldsymbol{\theta}$ identifiable in the whole parameter space? Answer Yes or No. If the answer is Yes, prove it. If the answer is No, give the set of points where the parameter vector is not identifiable.

vi. Give a simple numerical example of two parameter vectors with different β values, yielding the same covariance matrix $\boldsymbol{\Sigma}$ of the observable data.

vii. Is the entire parameter $\boldsymbol{\theta}$ identifiable where $\phi_{12} \neq 0$? If the answer is Yes, prove it.

viii. Why is it reasonable to assert $\phi_{12} \neq 0$?

(c) Now suppose X is measured with error as well as being correlated with omitted variables.



i. Give the model equations in centered form. The centering can be invisible.

ii. What is the parameter vector $\boldsymbol{\theta}$ for this model?

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

iv. Calculate the covariance matrix $\boldsymbol{\Sigma}$ of the observable data vector.

v. Is the parameter β identifiable provided $\phi_{12} \neq 0$? Answer Yes or No. If the answer is Yes, prove it. If the answer is No, give a simple numerical example of two parameter vectors with different β values, yielding the same covariance matrix $\boldsymbol{\Sigma}$ of the observable data.

2. 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, $Var(X_i) = \phi$, $Var(e_i) = \omega$, $Var(\epsilon_{i,1}) = \psi_1$, $Var(\epsilon_{i,2}) = \psi_2$, 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 is the parameter vector $\boldsymbol{\theta}$ for this model?

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

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

(e) 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?

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

(g) 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 $\beta_2 = 0$.

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

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

(j) Recall that an *instrumental variable* is an observable variable that has non-zero covariance with the explanatory variable and zero covariance with the error term in the regression. Under what condition is Y_2 an instrumental variable for X?

(k) 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 $H_0 : \beta_1 = 0$? If this does not bother you, it should.

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

3. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_{i,1} &= \alpha_1 + \beta_1 X_{i,1} + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta_2 X_{i,2} + \epsilon_{i,2} \\ W_{i,1} &= \nu_1 + X_{i,1} + e_{i,1} \\ W_{i,2} &= \nu_2 + X_{i,2} + e_{i,2} \\ V_{i,1} &= \nu_3 + Y_{i,1} + e_{i,3} \\ V_{i,2} &= \nu_4 + Y_{i,2} + e_{i,4}, \end{aligned}$$

where $E(X_{i,j}) = \mu_j$, $e_{i,j}$ and $\epsilon_{i,j}$ are independent of one another and of $X_{i,j}$, $Var(e_{i,j}) = \omega_j$, $Var(\epsilon_{i,j}) = \psi_j$, and

$$V \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}.$$

In this model, $X_{i,1}$, $X_{i,2}$, $Y_{i,1}$ and $Y_{i,2}$ are latent variables, while $W_{i,1}$, $W_{i,2}$, $V_{i,1}$ and $V_{i,2}$ are observable.

(a) Make a path diagram for this model.

(b) Does this model fit the double measurement design?

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

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

(e) Show that β_1 and β_2 are identifiable provided $\phi_{12} \neq 0$.

(f) Are there any other points in the parameter space where β_1 is identifiable? β_2?

(g) Give one testable equality constraint (a statement about the σ_{ij} quantities) that is implied by the model. Is it still true with $\phi_{12} = 0$? $\beta_1 = 0$? $\beta_2 = 0$?

(h) Suppose you wanted to estimate β_1. Suggest a *statistic* (function of the sample data) to serve as an estimator.

(i) Is your estimator consistent? Under what circumstances? You don't have to prove anything in detail.

(j) If the primary interest is in β_1, do we really need the response variable $Y_{i,2}$?

(k) Does any variable in this model qualify as an instrumental variable?

4. In this problem, W_i and the Y_{ij} variables are observable, and X_i is latent. The response variable of primary interest is $Y_{i,1}$, while $Y_{i,2}$ and $Y_{i,3}$ are included to help with identifiability. The point of the question is that the error terms need not all be independent for this to work.

   Independently for $i = 1, \ldots, n$,

$$\begin{aligned} Y_{i,1} &= \beta_{0,1} + \beta_{1,1} X_i + \epsilon_{i,1} \\ Y_{i,2} &= \beta_{0,2} + \beta_{1,2} X_i + \epsilon_{i,2} \\ Y_{i,3} &= \beta_{0,3} + \beta_{1,3} X_i + \epsilon_{i,3} \\ W_i &= X_i + e_i \end{aligned}$$

   where

   - $X_i \sim N(\mu_x, \phi)$ is a latent variable
   - $e_i \sim N(0, \omega)$
   - $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \epsilon_{i,2}, \epsilon_{i,3})^\top$
   - X_i, e_i and $\boldsymbol{\epsilon}_i$ are independent of one another
   - $\boldsymbol{\epsilon}_i$ is multivariate normal with mean zero and covariance matrix

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{1,1} & \psi_{1,2} & 0 \\ \psi_{1,2} & \psi_{2,2} & \psi_{2,3} \\ 0 & \psi_{2,3} & \psi_{3,3} \end{bmatrix}.$$

   (a) Make a path diagram for this model

   (b) What is the parameter vector $\boldsymbol{\theta}$ for this model?

   (c) How many moment structure equations are there? You do not have to say what they are; just give a number. Don't forget the means.

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

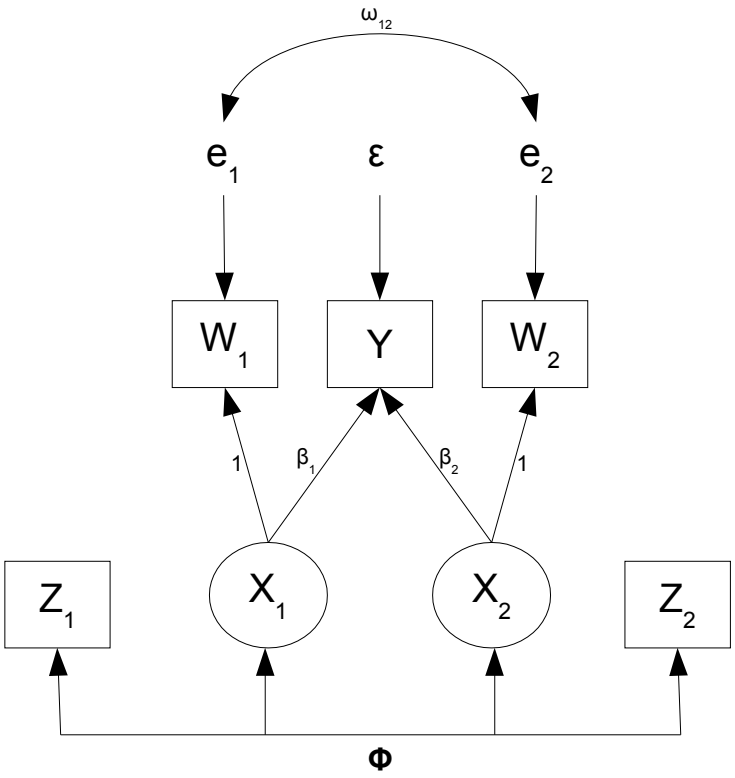
   (e) Calculate the variance-covariance matrix of the observable variables. Remember that some covariances between errors are non-zero. Show your work.

   (f) Solving the complete set of moment structure equations can be done[2] but it's a big chore. The primary interest is in the parameter $\beta_{1,1}$. Show that just this parameter is identifiable.

   (g) Does any variable in this model qualify as an instrumental variable?

[2] Even the intercepts are identifiable from the mean vector $\boldsymbol{\mu}$, because there is no measurement bias term in this model. That's unrealistic, of course.

5. Here is a model with two instrumental variables. Note that in this case there are omitted variables that affect the observable versions of both explanatory variables, so that the measurement error terms are correlated. This is usually poison.



(a) Give the model equations in centered form. The centering can be invisible.

(b) What is the parameter vector $\boldsymbol{\theta}$ for this model?

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

(d) Calculate the covariance matrix $\boldsymbol{\Sigma}$ of the observable data vector. When I did it I used the order of observable variables in the path diagram, but it will be easier to check identifiability if you write the observable data vector as $\mathbf{D} = (W_1, W_2, Z_1, Z_2, Y)^\top$.

(e) Are the parameters β_1 and β_2 identifiable? Answer Yes or No. If the answer is Yes, prove it. You don't have to finish solving for β_1 and β_2. Just give two linear equations in β_1 and β_2 as well as a number of σ_{ij} quantities. Presumably it's possible to solve two linear equations in two unknowns.

6. Question 8 (part of the SAS assignment) will use the *Longitudinal IQ Data*. IQ is short for "Intelligence Quotient," and IQ tests are attempts to measure intelligence. A score of 100 is considered average, while scores above 100 are above average and scores below 100 are below average. Most IQ tests have many sub-parts, including vocabulary tests, math tests, logical puzzles, tests of spatial reasoning, and so on. What the better tests probably succeed in doing is to measure one *kind* of intelligence — potential for doing well in school. Of course, they measure it with error.

   In the Longitudinal IQ Data, the IQs of adopted children were measured at ages 2, 4, 8 and 13. The birth mother's IQ was assessed at the time of adoption, and the adoptive mother's education (in years) was also recorded. The variables are

   - Adoptive mother's education
   - Birth mother's IQ
   - IQ at age 2
   - IQ at age 4
   - IQ at age 8
   - IQ at age 13

   In our dreams, we wish for a regression model in which the explanatory variables are adoptive mother's actual education (a latent variable), birth mother's true IQ (also latent), and child's IQ at ages 2, 4, 8 and 13 — all latent. Well, adoptive mother's education has only one measurement and no convincing instrumental variables, so we'll reluctantly set it aside for now.

   (a) To show you know what's going on, write down a regression model for just the IQ part of the data. My model has 5 latent variables and 5 observable variables. Give all the details. It has been verified many times that IQ scores have a normal distribution, so for once the normal distribution assumption is very reasonable.

   (b) As usual, set the intercepts and expected values aside. Calculate the covariance matrix in terms of the model parameters.

   (c) Does the model pass the test of the parameter count rule? Give the numbers.

   (d) To get out of this mess, we re-parameterize, combining the variance of ϵ and the variance of e into a single parameter for each response variable. This is equivalent to adopting a model with no measurement error in the response variables. So now we have a model that has one explanatory variable measured with error, and 4 response variables measured without error. Write the covariance matrix for this model, which you can mostly just copy from your earlier work.

   (e) Show that the parameters of your model (anyway, those appearing in the covariance matrix) are identifiable. What do you need to assume? What hypotheses would you test about single σ_{ij} quantities to verify this?

   (f) How many degrees of freedom should there be in the likelihood ratio test for model fit? The answer is a number.

(g) Suppose you want to test whether all the regression coefficients are equal, using a likelihood ratio test.

i. What are the degrees of freedom for this test?

ii. If you reject H_0, what will you conclude about how the birth mother's IQ is related to the child's IQ at various ages?

7. We have some unfinished business from the pig study of Assignment 6.

(a) You tested for correlated measurement error within questionnaires with two separate Z tests. It was pretty convincing, but 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.

(b) The double measurement design allows the measurement error covariance matrices $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ 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.

8. The longitudinal IQ data described in Question 6 are given in the file `origIQ.data.txt`. These data are taken from *The Statistical Sleuth* by F. Ramsey and D. Schafer, and are reproduced without permission. There is a link on the course web page in case the one in this document does not work. Note there are $n = 62$ cases, so 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 variables, including adoptive mother's education.

   (b) How are the `proc corr` results helpful in justifying your identifiability conditions from the Question 6?

   (c) Remember your model that has one explanatory variable measured with error, and 4 response variables measured without error? We'll call this the *full model*. Please fit the full model.

   (d) Sticking strictly to the $\alpha = 0.05$ significance level, does the full model fit the data adequately? Answer Yes or No, and give a value of G^2, the degrees of freedom and the p-value. These numbers are all directly on your printout. Do the degrees of freedom agree with your answer to Question 6f?

   (e) Now fit the reduced model in which all the regression coefficients are equal. Using a calculator (or `proc IML` if you want to), calculate the likelihood ratio test comparing the full and reduced models. Obtain G^2, a number.

   (f) What are the degrees of freedom for this test? Compare your answer to Question 6(g)i.

   (g) Using this table of critical values, do you reject H_0 at $\alpha = 0.05$? Answer Yes or No. Does birth mother's IQ seem to affect her child's IQ to the same degree at different ages?

```
> df = 1:8
> CriticalValue = qchisq(0.95,df)
> round(rbind(df,CriticalValue),3)
                [,1]  [,2]  [,3]  [,4]  [,5]   [,6]   [,7]   [,8]
df             1.000 2.000 3.000 4.000  5.00  6.000  7.000  8.000
CriticalValue 3.841 5.991 7.815 9.488 11.07 12.592 14.067 15.507
```

   To be continued ...

Bring your log files and your output files 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.**