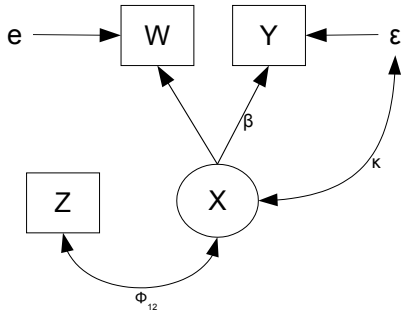
STA 431s17 Assignment Eight[1]

The first three questions of this assignment are about how instrumental variables can help with measurement error and omitted variables at the same time; see Lecture slide set 13. When instrumental variables are not available, sometimes extra response variables allow us to purchase identifiability. Questions 4 through 7 explore this possibility.

The non-computer questions on this assignment are for practice, and will not be handed in. For the SAS part of this assignment (Question 7) 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. In the model pictured below, the explanatory variable X is measured with error as well as being correlated with omitted variables. Z is an instrumental variable.



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

   (b) Guided by the symbols on the path diagram, provide notation for the variances and covariances of the error terms and exogenous variables.

   (c) Let $\boldsymbol{\theta}$ denote the vector of parameters you have witten down so far. These are the parameters that will appear 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. (Notice that we are only trying to identify the parameters in $\boldsymbol{\theta}$, which is a function of the full parameter vector. The full parameter vector has intercepts and unknown probability distributions.)

   (e) Calculate the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{D_i}$, a single observable data vector.

   (f) 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 $\boldsymbol{\Sigma}$.

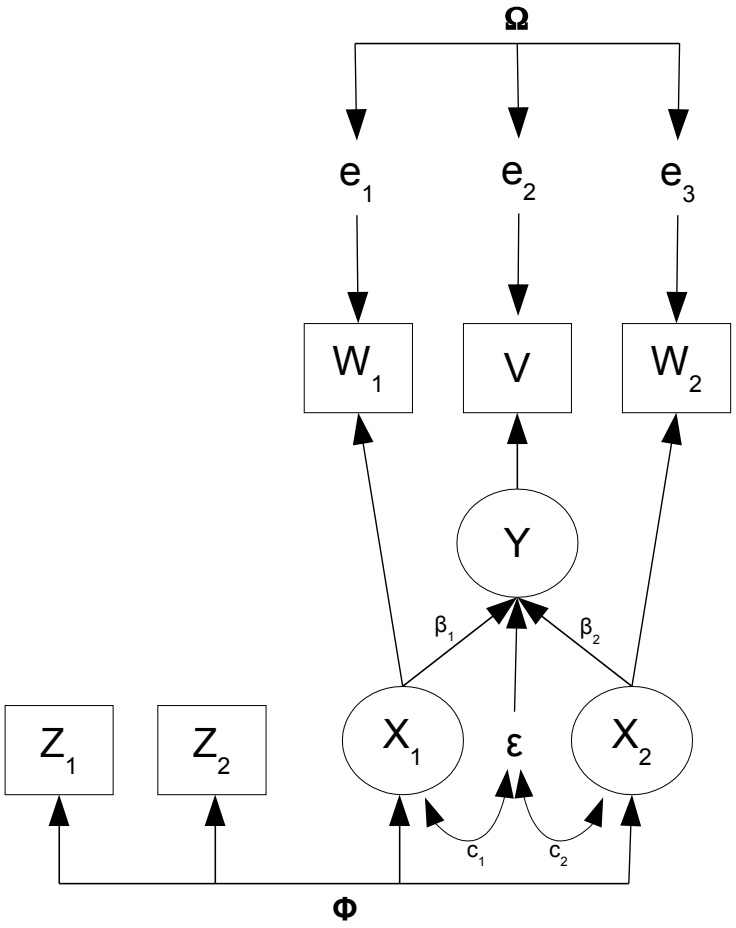
(g) Why is is reasonable to assume $\phi_{12} \neq 0$?

(h) Now let's make the model more realistic and scary. The response variable is measured with error, so $V = Y + e_2$. Furthermore, because of omitted variables, all the error terms might be correlated with one another and with X.

i. Do your best to make a path diagram of the new model. You need not write symbols on the curved double-headed arrows you have added.

ii. Show that β is still identifiable.

2. Here is a model with two explanatory variables and two instrumental variables. The path diagram looks busy, but it has features that make sense once you think about them. The instrumental and explanatory variables have covariance matrix $\mathbf{\Phi} = [\phi_{ij}]$, so that for example $Var(X_1) = \phi_{33}$. No doubt there are omitted explanatory variables that are correlated with X_1 and X_2, and affect Y. That is the source of $c_1 = Cov(X_1, \epsilon)$ and $c_2 = Cov(X_2, \epsilon)$. The variables in the latent regression model are all measured (once) with error. Because of omitted variables in the measurement process, the measurement errors are correlated, with 3×3 covariance matrix $\mathbf{\Omega} = [\omega_{ij}]$.



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

(b) How many parameters appear in the covariance matrix of the observable data? Scanning from the bottom, I get 10+2+2+6=20.

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

(d) The next step would be to calculate the covariance matrix $\boldsymbol{\Sigma}$ of the observable data vector, but that's a big job. To save work and also to reveal the essential features of the problem, please just calculate $cov\left((Z_1, Z_2)^\top, (W_1, W_2, V)^\top\right)$.

(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. You can stop once you have two linear equations in two unknowns, where the coefficients are σ_{ij} quantities. Presumably it's possible to solve two linear equations in two unknowns. To prove identifiability, you don't have to actually recover the parameters from the covariance matrix. All you have to do is show it can be done. In $\boldsymbol{\Sigma}$, please maintain the order Z_1, Z_2, W_1, W_2, V so we will have the same answer.

3. Here is a matrix version of instrumental variables. Independently for $i = 1, \ldots, n$, the centered model equations are

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

The random vectors $\mathbf{X}_i$ and $\mathbf{Y}_i$ are latent, while $\mathbf{W}_i$ and $\mathbf{V}_i$ are observable. In addition, there is a vector of observable instrumental variables $\mathbf{Z}_i$. The random vectors $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $p \times 1$, while $\mathbf{Y}_i$ is $q \times 1$. This determines the sizes of all the matrices. The variances and covariances are as follows: $cov(\mathbf{X}_i) = \boldsymbol{\Phi}_x$, $cov(\mathbf{Z}_i) = \boldsymbol{\Phi}_z$, $cov(\mathbf{Z}_i, \mathbf{X}_i) = \boldsymbol{\Phi}_{zx}$, $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$, $cov(\mathbf{X}_i, \boldsymbol{\epsilon}_i) = \mathbf{C}$, and $cov\begin{pmatrix} \mathbf{e}_{i,1} \\ \mathbf{e}_{i,2} \end{pmatrix} = \boldsymbol{\Omega}$. All variance-covariance matrices are positive definite (why not), and in addition, the $p \times p$ matrix of covariances $\boldsymbol{\Phi}_{zx}$ has an inverse. Covariances that are not specified are zero; in particular, the instrumental variables have zero covariance with the error terms.

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

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

where $cov(\mathbf{Z}_i, \mathbf{W}_i) = \boldsymbol{\Sigma}_{12}$, and so on.

(a) Give the dimensions (number of rows and number of columns) of the following matrices: $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Sigma}_{23}$.

(b) This problem fails the test of the Parameter Count Rule, though you are not required to show it. Fortunately, all we care about is $\boldsymbol{\beta}$. Doing as little work as possible, prove that $\boldsymbol{\beta}$ is identifiable by showing how it can be recovered from the $\boldsymbol{\Sigma}_{ij}$ matrices.

(c) Give the formula for an estimator of $\boldsymbol{\beta}$ and show that it is consistent.

4. 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, $Var(X_i) = \phi$, $Var(e_i) = \omega$, $Var(\epsilon_{i,1}) = \psi_1$, $Var(\epsilon_{i,2}) = \psi_2$, 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 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?

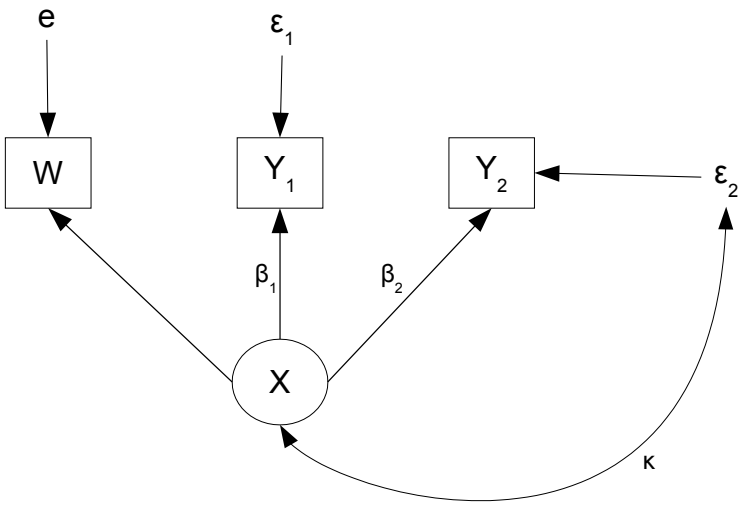
(n) How do you know that your Method of Moments estimator is also the Maximum Likelihood estimator (assuming normality)?

(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* constraints on the variances and covariances. Give at least one example.

5. In the model of Question 4, 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 $\boldsymbol{\Sigma}$) 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 4?

(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 4, 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?

(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 5e, what null hypothesis would you test about the σ_{ij} quantities to test $H_0 : \beta_1 = 0$?

ii. If you failed to reject the null hypothesis in Question 5e, could you still test $H_0 : \beta_1 = 0$? What is the test on σ_{ij} quantities in this case?

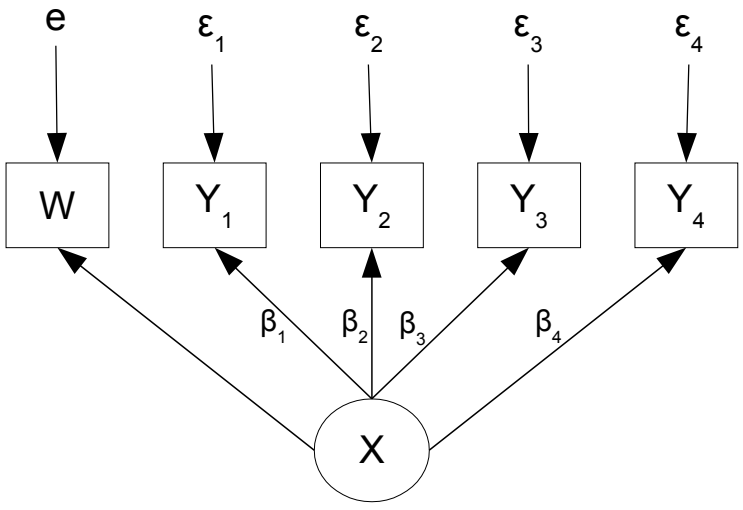
iii. If you rejected $H_0 : \beta_1 = 0$, naturally you would want to state whether β_1 is positive or negative. Is this possible?

6. Question 7 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. 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

   Here is a path diagram for just the IQ part of the data.



X is birth mother's true academic ability, W is birth mother's measured IQ, Y_1 is child's measured IQ at age 2, and so on. Of course child's ability is a latent variable too, but we've re-parameterized, making the response variables appear to be observable. This model is far from perfect, but we'll go with it for now.

(a) Write the model equations in scalar form.

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

(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 vector is identifiable except on a set of volume zero in the parameter space. Show how β_2 can be recovered from σ_{ij} quantities.

(f) If $\beta_1 = 0$, is it identifiable?

(g) How many βs need to be non-zero for all the parameters to be identifiable?

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

(i) Suppose you want to test whether all four regression coefficients are equal, using a likelihood ratio test. What are the degrees of freedom for this test?

7. 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 used without permission.

(a) Start by reading the data. There are $n = 62$ cases, so please verify that you are reading the correct number of cases. Now run `proc corr` to produce a correlation matrix of all the variables (with tests), including adoptive mother's education.

(b) Do the test on the correlations support omitting adoptive mother's education from the model?

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

(d) Please fit the model of Question 6. We'll call this the *full model*.

(e) Sticking strictly to the $\alpha = 0.05$ significance level, does the full model fit the data adequately? Answer Yes or No[2], 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 6h?

[2]My p-value of 0.0707 is too close to 0.05 for comfort. This model is hard to defend. It says the only reason that kids' scores at different ages are correlated is the mother's academic potential as an adult. Just for starters, these children do technically have fathers. Also, the model fails to distinguish between genotype and phenotype, and it has other weaknesses.

(f) 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.

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

(h) 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
```

(i) Just one more thing. Please go back to the `proc calis` run where you fit the full model, and use the `simtests` statement to do a Wald test of $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$. Do you reach the same conclusion? SAS makes it easy to do Wald tests and hard to do likelihood ratio tests, but likelihood ratio tests perform better for relatively small samples like this, so we'll believe the likelihood ratio test.

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.** If you see a reference to the Moore-Penrose inverse, you've made a coding error.