STA 431s13 Assignment Six[1]

For the SAS question, please bring your log and list files to the quiz. Do not write anything on the printouts except your name and student number. The other questions are just practice for the quiz on Friday March 1st, and are not to be handed in.

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

$$\begin{aligned} W_{i,1} &= X_{i,1} + e_{i,1} \\ W_{i,2} &= X_{i,2} + e_{i,2} \\ Y_{i,1} &= \beta_1 X_{i,1} + \epsilon_{i,1} \\ Y_{i,2} &= \beta_2 X_{i,2} + \epsilon_{i,2} \\ Y_{i,3} &= \beta_3 X_{i,1} + \beta_4 X_{i,2} + \epsilon_{i,3} \end{aligned}$$

   where

   - The $X_{i,j}$ variables are latent, while the $W_{i,j}$ and $Y_{i,j}$ variables are observable.
   - $e_{i,1} \sim N(0, \omega_1)$ and $e_{i,2} \sim N(0, \omega_2)$.
   - $\epsilon_{i,j} \sim N(0, \psi_j)$ for $j = 1, 2, 3$.
   - $e_{i,j}$ and $\epsilon_{i,j}$ are independent of each other and of $X_{i,j}$.
   - $X_{i,j}$ have expected value zero and

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

   Denote the vector of observable data by $\mathbf{D}_i = (W_{i,1}, W_{i,2}, Y_{i,1}, Y_{i,2}, Y_{i,3})'$, with $V(\mathbf{D}_i) = \boldsymbol{\Sigma} = [\sigma_{ij}]$.

   Among other things, this question illustrates how the search for identifiability can be supported by exploratory data analysis. Hypotheses about *single* covariances, like $H_0 : \sigma_{ij} = 0$ can be tested without effort by looking at tests of the corresponding correlations. These tests are produced automatically by `proc corr`.

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

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

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

   (d) The parameter ϕ_{12} is identifiable. How?

[1] Copyright information is at the end of the last page.

(e) Suppose $\beta_1 = 0$. Why is the parameter β_1 identifiable? Of course the same applies to β_2.

(f) But the idea here is that Y_1 and Y_2 are instrumental variables, so that $\beta_1 \neq 0$ and $\beta_2 \neq 0$. What hypotheses about *single* covariances would you test to verify this?

(g) From this point on, suppose we have verified $\beta_1 \neq 0$ and $\beta_2 \neq 0$. Under what circumstances (that is, where in the parameter space) can the parameters β_1 and β_2 be easily identified?

(h) What hypotheses about *single* covariances would you test to persuade yourself that this is okay?

(i) Assuming the last step worked out well, give a formula for β_1 in term of σ_{ij} values.

(j) Suppose you were sure $\phi_{12} \neq 0$, but you were not so sure about normality so you were uncomfortable with maximum likelihood estimation. Suggest a nice estimator of β_2. Why are you sure it is consistent? Note that even if you were interested in the MLE, this estimate would be an excellent starting value.

(k) Suppose your test for $\phi_{12} = 0$ did not reject the null hypothesis, so dividing by σ_{12} makes you uncomfortable. Show that even if $\phi_{12} = 0$, there is another way to identify β_1. What assumption to you have to make (that is, where in the parameter space does the true parameter vector have to be) for this to work? How would you test it?

(l) How could you identify β_2 if $\phi_{12} = 0$?

(m) In question 1j, you gave an estimator for β_2 that is consistent in most of the parameter space. Based on your answer to the preceding question, give a second estimator for β_2 that is consistent in most of the parameter space. It should be geometrically obvious that except for a set of volume zero in the parameter space, *both* estimators are consistent.

(n) Assuming β_1 and β_2 are identifiable one way or the other, now we seek to identify ϕ_{11} and ϕ_{22}. How can this be done? Give the formulas. Also, give a consistent estimator of ϕ_{22} that is not the MLE. Why are you sure it's consistent?

(o) Since Y_1 and Y_2 are instrumental variables, primary interest is in β_3 and β_4, the coefficients linking Y_3 to X_1 and X_2. If our efforts so far have been successful (which they are, except on a set of volume zero in the parameter space), then β_3 and β_4 can be identified as the solution to two linear equations in two unknowns. Write these equations *in matrix form*.

(p) What condition on the ϕ_{ij} values ensure a unique solution to the two equations in two unknowns? Is this a lot to ask?

(q) Now let's back up, and admit that the identification of β_3 and β_4 is really the whole point, since they are the parameters of interest. We have seen that ϕ_{12} is always identifiable. If $\phi_{12} \neq 0$, it can be used to identify β_1 and β_2, and they can be used to identify ϕ_{11} and ϕ_{22}. Then β_3 and β_4 can be identified by solving the

two equations in two unknowns. Now suppose that $\phi_{12} = 0$. In this case β_3 and β_4 can be identified without knowing the values of ϕ_{11} and ϕ_{22}, provided β_1 and β_2 are non-zero. Show how this can be done.

(r) Assuming that the parameters appearing in the covariances of $\mathbf{\Sigma}$ are identifiable, the additional 5 parameters (whch appear only in the variances) may be identified by subtraction. So we see that except on a set of volume zero in the parameter space, all the parameters are identifiable. In that region, how many equality constraints should the model impose on the covariance matrix? Use your answer to Question 1b.

(s) To see what the equality constraints are, note that earlier parts of this question point to two ways of identifying β_1 and two ways of identifying β_2. There are also two simple ways to identify ϕ_{12}. So write down the three constraints. Multiply through by the denominators.

(t) Now you have three equalities involving products of σ_{ij} terms. For each one, use your covariance matrix to write both sides in terms of the model parameters. For each equality, does it hold everywhere in the parameter space, or are there some points in the parameter space where it does not hold? If there are points in the parameter space where an equality does not hold, state the set explicitly.

(u) The idea here is that the three degrees of freedom in the likelihood ratio test of model fit correspond to three equalities involving the covariances, and those equalities are directly testable without the normality assumption[2] required by the likelihood ratio test. State the null hypothesis (there's just one) in terms of the σ_{ij} quantities.

(v) If the null hypothesis were rejected, what would you conclude about the model?

(w) In ordinary multivariate regression (which has more than one response variable), it is standard to assume the error terms for the response variable may have non-zero covariance. Suppose, then, that $Cov(\epsilon_{i,1}, \epsilon_{i,2}) = \psi_{12}$. How would this change the covariance matrix?

(x) Always remembering that β_1 and β_2 are non-zero, suppose that $\phi_{12} = 0$. Is ψ_{12} identifiable? What if $\phi_{12} \neq 0$?

(y) Well, what if there were non-zero covariances ψ_{13} and ψ_{23} as well? What does the parameter count rule tell you?

(z) Again by the parameter count rule, $\phi_{12} \neq 0$ is absolutely necessary to identify the entire parameter if all three ψ_{ij} are added to the model. Why? In this case, are ψ_{13} and ψ_{23} identifiable? Why or why not?

[2]It's true that I have not told you how to do this yet, but it's not hard.

2. Question 3 (the SAS part of this 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. And 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 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. 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 $sigma_{ij}$ 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?

3. The longitudinal IQ data are given in the file `origIQ.data`. 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. The `proc corr` procedure is illustrated in SAS Example One.

(b) How are the `proc corr` results helpful in justifying your identifiability calculations from the last question?

(c) Remember your model 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 2f?

(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 2(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
```

(h) That was interesting, but now let's bring in adoptive mother's education. We wonder whether, controlling for birth mother's true IQ, adoptive mother's true education is related to the child's true IQ. Write the model equations, using notation similar to Question 1. There are now *two* regression coefficients for each response variable. Don't bother with the intercepts.

(i) Calculate the covariance matrix in terms of the model parameters. Does this model pass the test of the parameter count rule?

(j) There is still hope. Your model has a term ϕ_{12}, representing the covariance between birth mother's true IQ and adoptive mother's true education. But unless the adoption agency acted in a very peculiar way, there is no reason these variables should be related. Furthermore, it's testable without actually fitting the model under consideration. Locate the test on your printout (it's there) and give the p-value.

(k) This is exploratory data analysis, so let's tentatively accept the (null) hypothesis $\phi_{12} = 0$. Under this assumption, your covariance matrix simplifies quite a bit. Either re-write it, or else circle the terms with ϕ_{12}, to remind yourself that they equal zero.

(l) The regression coefficients linking adoptive mother's education to child's IQ at various ages are now identifiable, meaning they are identifiable at points in the parameter space where $\phi_{12} = 0$. Recover one of them from the covariance matrix just to show you can do it.

(m) Under the null hypothesis that adoptive mother's true education has no effect on child's IQ at any age, four covariances in $\boldsymbol{\Sigma}$ should be zero (assuming $\phi_{12} = 0$, of course). Which ones are they?

(n) Give the p-values, numbers from your printout. What do you conclude? Do these data support a link between adoptive mother's education and child's IQ?

This story could be continued a bit more, but that's pretty good. The lesson is that valid inference about a latent variable may be possible even when the model parameters cannot be estimated.
