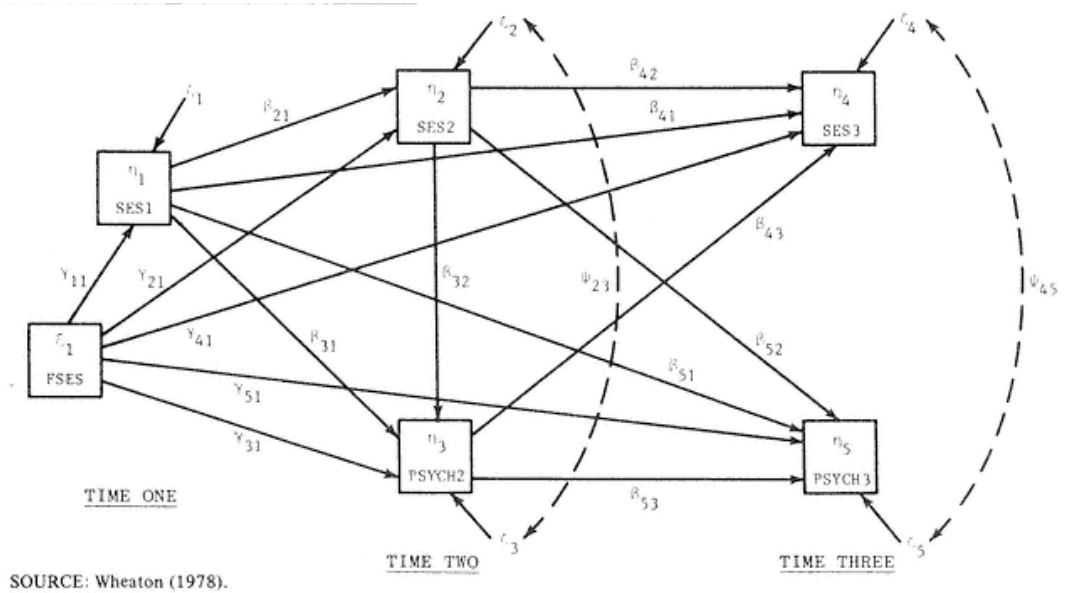
STA 431 Assignment 11

Do this assignment in preparation for the quiz on Friday, April 3d. Bring your log and list file for Question 5. The other questions are practice for the quiz, and are not to be handed in.

1. Here is a path diagram from a Sociology study.



SOURCE: Wheaton (1978).
NOTE: FSES = father's socioeconomic status; SES1 = socioeconomic status at time 1; SES2 = socioeconomic status at time 2; SES3 = socioeconomic status at time 3; PSYCH2 = number of psychological sysmptoms at time 2; PSYCH3 = number of psychological symptoms at time 3.

This is the initial model. The authors eliminated three arrows to make it identified, but you can do better. Eliminate just one arrow to make the model identified, and cite the rule(s) you are using.

2. The latent part of the LISREL structural equation model is

$$\boldsymbol{\eta} = \boldsymbol{\beta}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta},$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ are independent, with $V(\boldsymbol{\xi}) = \boldsymbol{\Phi}$ and $V(\boldsymbol{\zeta}) = \boldsymbol{\Psi}$. Derive

$$\boldsymbol{\Sigma}_0 = V\left[\begin{array}{c} \boldsymbol{\xi} \\ \boldsymbol{\eta} \end{array}\right].$$

Show your work. The answer is a partitioned matrix (a matrix of matrices).

3. Consider the confirmatory factor analysis model

$$\begin{aligned} X_1 &= F_1 + e_1 \\ X_2 &= \lambda_2 F_1 + e_2 \\ X_3 &= \lambda_3 F_1 + e_3 \\ X_4 &= F_2 + e_4 \\ X_5 &= \lambda_5 F_2 + e_5, \end{aligned}$$

where $e_1, \ldots, e_5$ are independent of one another and of F_1 and F_2, all expected values are zero, $V(e_i) = \psi_i$ for $i = 1, \ldots, 5$,

$$V \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{1,2} & \phi_{2,2} \end{bmatrix},$$

λ_2, λ_3 and λ_5 are nonzero constants, and all the distributions are normal.

(a) Give the covariance matrix of the observable variables. Show your work.

(b) What is the parameter vector θ for this model?

(c) Show that the model is identified *given one additional condition that has not been stated*. Find the condition, state it clearly and *circle it*. It is easiest to start proving that the model is identified, and find out what else you need along the way.

4. In lecture, we saw that negative variance estimates can be caused by correlated error terms. That is, the error terms are correlated in reality, but not in the model. This is another example. Let

$$\begin{aligned} X_1 &= F + e_1 \\ X_2 &= \lambda_2 F + e_2 \\ X_3 &= \lambda_3 F + e_3, \end{aligned}$$

where all expected values are zero, e_1, e_2 and e_3 are independent of F, $Var(F) = \phi$,

$$V \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} \psi_{1,1} & \psi_{1,2} & 0 \\ \psi_{1,2} & \psi_{2,2} & 0 \\ 0 & 0 & \psi_{3,3} \end{bmatrix},$$

λ_2 and λ_3 are nonzero constants, and all the distributions are normal. The observed variables are X_1, X_2 and X_3. In *Model One*, we will assume that the e_1 and e_2 are independent; that is, $\psi_{1,2} = 0$. In *Model Two*, e_1 and e_2 are not independent, and $\psi_{1,2} > 0$.

(a) Draw a path diagram of Model One. Now using a dotted line, put in one curved double-headed arrow representing the difference between Model One and Model Two.

(b) How do you know that the assumption $\psi_{1,2} = 0$ is necessary for the model to be identified? You do not need to do elaborate calculations; just cite a well-known rule.

(c) Assuming Model One, calculate the variance-covariance matrix of the observed variables. Express this matrix $\boldsymbol{\Sigma} = [\sigma_{i,j}]$ as a function of the model parameters. Show your work.

(d) Obtain explicit solutions of the identifying equations under Model One. Show your work. For later reference, you are writing $\boldsymbol{\theta} = \sigma^{-1}(\boldsymbol{\Sigma})$.

(e) Recall the invariance principle of maximum likelihood estimation, which says that the MLE of a function of the parameter is that same function of the MLE. Denoting the maximum likelihood estimate of $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}} = [\widehat{\sigma}_{i,j}]$, give an explicit formula for $\widehat{\psi}_{1,1}$. (We are still assuming Model One.) **Circle your answer.**

(f) By the Law of Large Numbers and continuous mapping, $\sigma^{-1}(\widehat{\boldsymbol{\Sigma}}) \to \sigma^{-1}(\boldsymbol{\Sigma})$, whether the model is right or not. Suppose that Model Two is correct, but you calculate the MLE under Model One (because it's really all you can do). What is the large-sample target of the variance estimate $\widehat{\psi}_{1,1}$? **Circle your answer.**

(g) What is the asymptotic bias of $\widehat{\psi}_{1,1}$? You don't need to know the formal definition of asymptotic bias to answer this question.

(h) Under what condition will the variance estimate $\widehat{\psi}_{1,1}$ be negative for large samples? Your answer is an inequality in the parameters of Model Two.

5. The file `peru.data` (see link on the course web page in case this one does not work) contains data from a study examining the long-term effects of change in environment on heart health. Subjects in this study were Aboriginals who had migrated from a very primitive environment, high in the Andes mountains of Peru, into the mainstream of Peruvian society, at a much lower altitude.

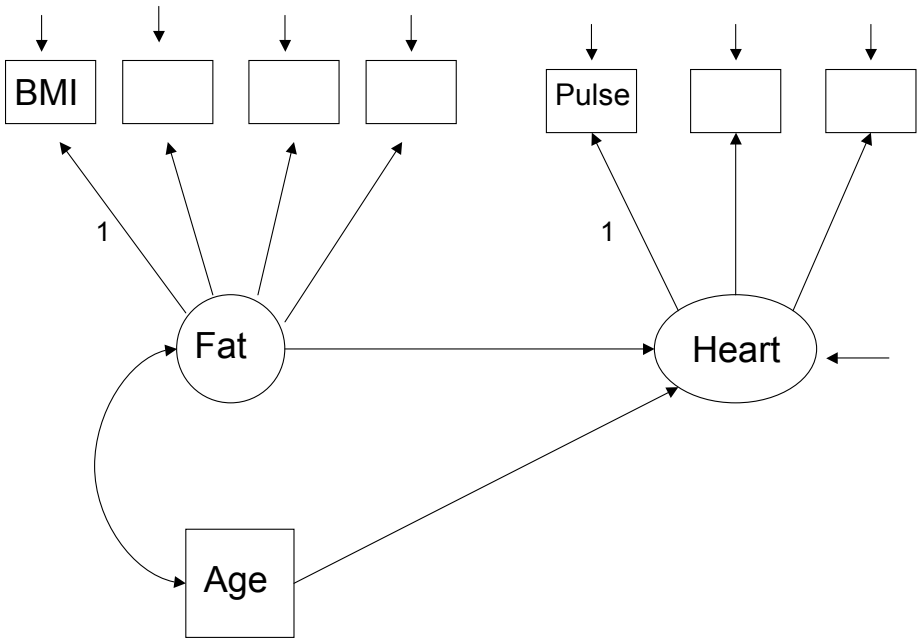
   All the data are for males over 21 who were born at high altitude and whose parents were born at high altitude. The variables in the file are:

   - Age in years
   - Years since migration (We will not be using this one.)
   - Weight in kilograms
   - Height in millimeters
   - Chin skin fold in millimeters
   - Forearm skin fold in millimeters
   - Calf skin fold in millimeters
   - Pulse rate in beats per minute

- Systolic blood pressure
- Diastolic blood pressure

The skin fold measures were taken as a general indiction of obesity, which tends to be a problem in this population.

Below is a *partial* path diagram of a Structural Equation Model for these data. Notice that age is assumed to be measured without error. Given where these guys come from, this is a little questionable.



(a) Start by filling in the blanks in the path diagram.

(b) Classify the ten variables (the ones that are not error terms) as either latent or observed, and as either endogenous or exogenous.

(c) Write the model equations in scalar (non-matrix) form. Use LISREL-type notation. I have eight equations.

(d) What is the parameter vector $\boldsymbol{\theta}$ for your model? You should have 18 parameters.

(e) Prove that the model parameters can be identified from the covariance matrix. Note that the model has both latent variables *and* an observed exogenous variable.

(f) Use `proc calis` to fit your model by maximum likelihood. You may want to use this line in the data step:

```
bmi = weight/(height/1000)**2;
```

(g) What does the chisquare test indicate about the fit of the model? Choose one: The fit is

i. Okay

ii. Not Great

iii. Terrible

Support your conclusion by citing two numbers from the printout: the value of a test statistic, and a p-value.

(h) There seems to be an outlier, and maybe he is responsible for the poor fit. Try

```
proc plot; plot weight*height;
```

This person is hefty.

(i) Who is he? Try

```
proc univariate; var weight bmi;
```

Part of the output from `proc univariate` is "Extreme Observations," which gives you the 5 biggest and 5 smallest values of a variable. And "Obs" is observation number, numbered from the top of the file. Look in `peru.data`. Have you found him?

(j) Okay, now create a new data file called something like `peru2.data`, and edit it to eliminate the outlier. Run `proc calis` on these data. How is the chisquare test now?

(k) What is the MLE of the coefficient linking fatness to heart problems? The answer is a number from your printout.

(l) Using the usual significance level $\alpha = 0.05$ is there evidence that controlling for age, fatness is causing heart problems? Support your conclusion by citing the value of a test statistic from the printout. What is the null hypothesis, and what value of the test statistic would lead you to reject it?

(m) We are using large-sample tests, which are justified as $n \to \infty$. What is the sample size? Please comment.