STA 431s17 Final Computer Assignment[1]

The SAS part of the final exam will be based on the Arthritis data. The data are simulated, but engineered to reproduce sample statistics from the baseline period of an actual study. This way there are no issues of data ownership or copyright, and also the study design is better than the original. On the exam you will answer questions based on my SAS program and output. To prepare, you should fit some models yourself, familiarize yourself with the output and think about what it means. There is no doubt that you will be asked for plain-language conclusions, and you might as well be ready. Eman and I will answer questions during the exam, but only if the answers are very short. You are expected to be familiar with this material.

In a study of exercice and arthritis pain, rheumatoid arthritis patients were clinically assessed for disease severity by a physician. Disease severity was also estimated by X-rays (based on joint erosion) and a blood test (based on elevated ESR and C-reactive protein, rheumatoid factor and anti-citrullinated protein antibody). The doctor made the clinical assessment before seeing the X-ray and blood test results.

One week later, patients and their spouses came into the clinic again. They both filled out questionnaires and the patients had more tests. Pain was measured in two ways: self-report about pain during the preceding week, and an electroencephalograph (EEG, or brain wave) test. In the EEG test, electrodes on the patient's scalp measured electrical activity in the brain during a standard passive joint movement exam. Passive means the patient relaxes while a technician moves the joint gently through a moderate range of motion. In the absence of loud noises or emotionally arousing stimuli, general autonomic nervous system activation is a fairly dependable indication of subjective pain.

Exercise/physical activity level during the preceding week was measured in three ways: self-report, spouse's report, and by accelerometer, a motion detector/fitness tracker that the patient had been wearing during the past week.

One week after that, patients and their spouses came in again, and the same measurements of pain and exercise were collected a second time. Here are the observable variables, in the form of a `label` statement from the SAS `data` step. This is part of my SAS program, which will be displayed on the final exam. You might as well use it directly.

```
label clinical    = 'Disease severity based on clinical assessment'
      xray         = 'Disease severity based on x-ray'
      blood        = 'Disease severity based on blood test'
      selfpain1    = 'Self-reported pain at time one'
      EEG1         = 'Pain assessed from brain waves at time one'
      selfexer1    = 'Self-reported exercise/physical activity at time one'
      spouseexer1  = 'Spouse report of exercise/physical activity at time one'
      acceler1     = 'Accelerometer (fitness tracker) data at time one'
      selfpain2    = 'Self-reported pain at time two'
      EEG2         = 'Pain assessed from brain waves at time two'
      selfexer2    = 'Self-reported exercise/physical activity at time two'
      spouseexer2  = 'Spouse report of exercise/physical activity at time two'
      acceler2     = 'Accelerometer (fitness tracker) data at time two';
```

The raw data are in the form of a Microsoft Excel spreadsheet. They are available from

http://www.utstat.toronto.edu/∼brunner/data/legal/Arthritis1.xls .

My (surrogate) measurement model is routine, but I doubt that you would come up with exactly my latent variable model. Like all models it's debatable, but this is what will be on the final exam. Here is a path diagram, followed by a bit of explanation.
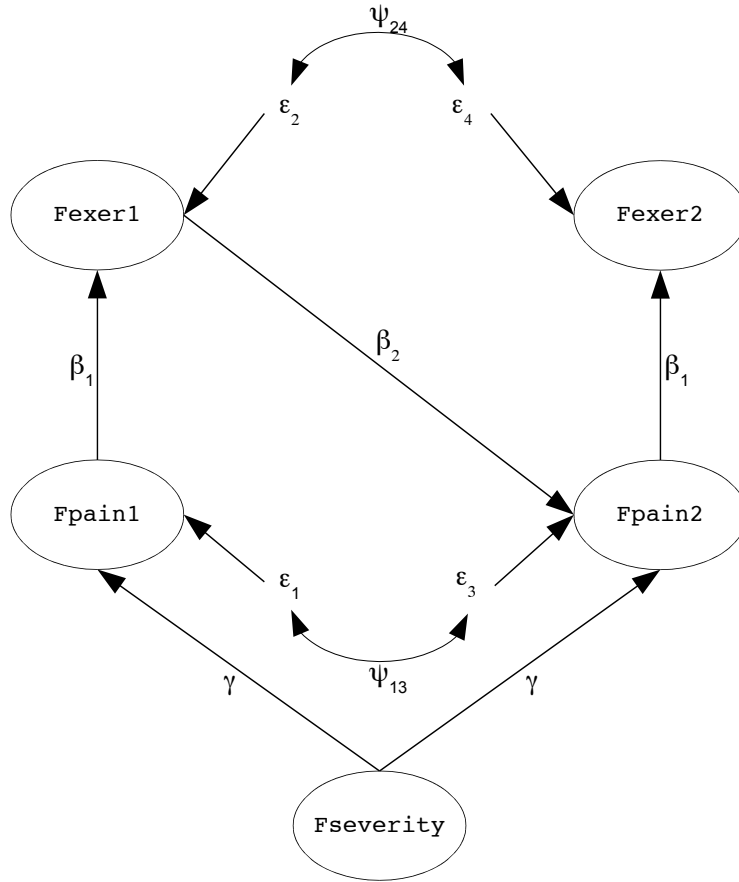
Rheumatoid arthritis is an auto-immune disorder with no known cure. It tends to get worse very gradually over time. This is why there's only one latent disease severity. This is a fairly stable system, so severity affects pain in the same way at both time periods. Similarly, pain affects exercise the same way at both time periods. Exercise at Time One may affect pain at Time Two; certainly the conventional wisdom is that it helps[2].

The curved arrows between error terms are what makes the model unusual. Pain really has momentum. Once it gets going it's harder to block, and it could be that pain at time one is directly contributing to pain at time two. But there are other things that would help produce a positive covariance between true pain at time one and true pain at time two, quite apart from disease severity. One omitted variable is the person's pain sensitivity, and there may be more. This is the reasoning behind the curved arrow connecting ϵ_1 and ϵ_3.

Exercise has momentum for reasons that are even more obvious. There's habit (established before the study began), social obligations to workout partners, New Year's resolutions, and just plain enjoyment of exercise (or the opposite). This explains the curved arrow between ϵ_2 and ϵ_4.

I hope you have been wondering about identifiability. If the curved arrows were replaced by straight arrows from Time One to Time Two, this model would satisfy the Acyclic Rule, with one variable in each set. The curved and straight arrows play the same role in the covariance matrix, and everything is okay. What would kill identifiability would be to have them both, because then they would be redundant.

---

[2]I think this is really interesting. If you run a `proc corr` (I will not do this on the exam), you will see negative correlations between the exercise measurements at Time One and the pain measurements at Time Two. This seems to support the conventional wisdom, but not so fast! The counter-argument is this. Because of disease severity, the more pain at Time One, the more pain at Time Two. And because exercise hurts when you have this disease, the more pain at Time One the less exercise at Time One. Therefore, the less exercise at Time One the more pain at Time Two, even if there is no direct link.