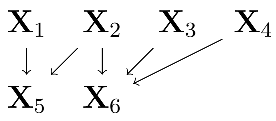
2023 STAT 528: Final

Due Thursday, March 9 at 12:30pm

Important instructions By submitting your final exam, you are certifying that this is your own work. You are not allowed to discuss this exam with anyone except the class instructor and the TA. You are not allowed to post questions on Canvas discussion boards. Instead, if you have questions, please email them to both the TA and the instructor. Please put “528 final” in the subject to make sure your emails are promptly given attention. The TA or the instructor will post general responses to your questions on Canvas.

1 Causal structure learning (30 points)

Consider the following graph and corresponding structural equation model



$$\begin{aligned}
&\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), \\
&\mathbf{X}_5 = 0.5\mathbf{X}_1 + 0.5\mathbf{X}_2 + \varepsilon_5 \\
&\mathbf{X}_6 = 0.5\mathbf{X}_2 + 0.5\mathbf{X}_3 + 0.5\mathbf{X}_4 + \varepsilon_6
\end{aligned} \tag{1}$$

Please answer the following questions.

1. Is the causal graph identifiable? Why or why not?

2. Draw $n = \{50, 100, 200, 400, 800\}$ samples of the variables $(\mathbf{X}_1, \ldots, \mathbf{X}_6)$. Use the GES algorithm with the BIC score to learn a graph and compare the estimate with the true graph. To compute a distance metric, compute the structural hamming distance (SHD) between the estimated and true skeletons. Repeat this for over 50 trials and report average SHD (averaged across the 50 trials) for all the sample sizes.

3. Repeat the previous part using the PC algorithm with $\alpha = 0.05$. How does the performance of the PC algorithm compare to GES?

2 Real data analysis (70 points)

You will analyze a data set from the National Institute of Mental Health Schizophrenia Collaborative Study on treatment-related changes in overall symptom severity in a sample of schizophrenic patients. Subjects were assigned to one of four treatments: placebo, chlorpromazine, fluphenazine, and thioridazine. In this version of the data set, the three non-placebo drug groups have been combined into one treatment group. Severity of schizophrenic symptomatology was tracked during the course of the study. Measurements were taken in Weeks 0, 1, 3, and 6. Conduct an analysis and write a report to address the following two questions:

Question 1: Researchers are interested to see if the treatment assignment significantly contributed to patient recovery over the course of 6 weeks. Is there a significant difference between the placebo and treatment groups at the end of the study?
Question 2: This data set has some missing values. Are you confident about your conclusion from Question 1 given the presence of missing data?

2.1 Data background

Severity of schizophrenic symptomatology was assessed across time using the seven-point Inpatient Multidi- mensional Psychiatric Scale (IMPS) Item 79, “Severity of Illness,” coded from 1 to 7 as follows:

1. not at all ill

2. borderline mentally ill

3. mildly ill

4. moderately ill
5. markedly ill
6. severely ill
7. among the most extremely ill

Patients were sometimes classified by two psychiatric raters (in terms of the severity as measured by this scale) and when these raters differed an average of the two scores was used for that patient at that time point. The file SCHIZREP.DAT.TXT, a subset of the full study data set, has the following five fields:

1. Field 1: patient ID
2. Field 2: IMPS79 (the 7-point measure of illness severity)
3. Field 3: week - from 0 (baseline) to 6
4. Field 4: treatment group (0 = placebo, 1 = drug)
5. Field 5: binary sex (0 = female, 1 = male)

2.2 Instructions for analysis

Some guidance and suggesstions are provided below. However, it is your responsibility as an analyst to decide on a way to analyze this data in order to address the research questions.

Make sure that you familiarize yourself with the data (observed and missing) using numerical summaries and graphical techniques before starting any analyses. You may recode treatment from numeric variable to a factor which is a good practice for categorical variables.

Please provide a justification for the model of your choice (why this particular model?), identify which estimate do you need to use to address the research questions, and decide on a strategy for handling missing data. It is suggested to use mixed effects models that include treatment group, time, treatment-time interaction as fixed effects and appropriate random effects, but you may consider different approaches as you see fit. If you use the suggested model, please also explain why this model makes sense to you. Explain your models and interpret estimates in sentences that psychiatric doctors can understand.

2.3 Instructions for writing the report

Please adhere to best practices for technical writing, and pay particular attention to presentation of graphics and tables.

Your final report needs to have the following sections (suggested space allocation is given in parentheses):

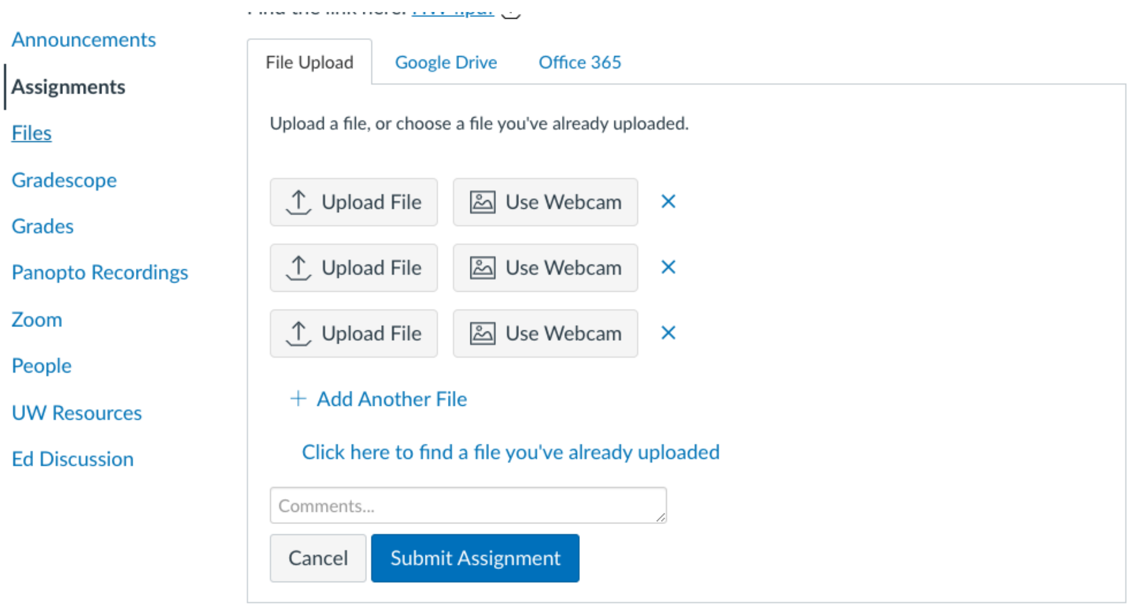
1. Introduction
2. Main Body
3. Summary/Conclusions
4. Appendix
5. References

The conclusion section should use the results that were obtained to answer the stated research question, and discuss difficulties and limitations of the analyses, if appropriate. You are welcome to provide a few key references on the methods used that will not count toward the six page total, but this is not required.

Please keep your report to six pages or less. You may include up to 10 tables and figures (e.g, four tables and six figures but not ten tables and ten figures) in your report. Do not include any plots, tables or results that you do not explicitly discuss. References do not count toward the page limit. It is acceptable to have some of the plots in an appendix that will not count toward your page limit (making sure they are clearly labeled/titled) as long as you summarize your observations about them in your paper, referencing their labels/titles so we know which plots you are referring to. Budget your time appropriately to account for writing and revisions, including getting your report to six pages or less.

Do not include any raw output (such as the results of summary(mymodel) or the messages from loading a library). Please keep all your final code and submit it so that we can reproduce your results, but do not include any code in your final report.

Proofread or skim your report to check that your figures, text, and sections appear as you expect them to, that your code is hidden, and so on. We recommend you do this periodically as you work on your analysis, not just at the end before submission. Please submit both your report and your code file. Don't compress them as a zip file. Instead, please you can click the button 'Add Another File' to upload multiple files at the same time, as illustrated by the following screen shot.



Upload a file, or choose a file you've already uploaded.

Figure 1: A screen shot on how to submit multiple files at Canvas.