Midterm (takehome) exam
UMN STAT 5511 (Fall 2024)

Charles R. Doss

Exam Instructions

Assigned: Monday, Nov 25, 2024 5:30pm.

Due: Exam must be handed in electronically (uploaded to Canvas) by the due date, Dec 3 at 11:59pm (Central time). Upload your exam as a *single pdf file.* Give yourself time to upload it so you do not miss the deadline. If you miss the deadline (even by 1 second), you may turn it in up to 24 hours late, but then there is a 15 point deduction. (The same timeline and 15 point deduction applies to hard copy exams handed in late.)

This is a takehome exam. You may not work (give or receive help) with any one else on this exam; that includes other students, as well as friends, family, colleagues, faculty, or anyone else. If you have questions you should contact the professor. You may reference all of your course materials (as well as outside textbooks or the internet). (However: for full credit you will generally need to give an answer that refers to what you learned specifically in class. If you use ideas or material from outside class, you will need to explain what you are doing, and they very often may not yield full credit.) You should not ask an Artificial Intelligence / large language model to solve the questions for you.

Note on academic honesty: statistics students have been caught cheating by working on takehome exams together. Any students caught cheating in such a manner will fail the course and have a mark put on their academic record, as a minimum punishment. Being caught cheating can lead to expulsion from the university (and loss of a student visa).

Questions

On this takehome exam, you will analyze three data sets. For the data sets, find `mid-data.rsav` on the course webpage. Load it into R by running `load("path/mid-data.rsav")` where "path/mid-data.rsav" is replaced by the full path on your hard drive to the file mid-data.rsav (the syntax for which is operating system dependent). The file contains two data objects: `dat1`, `dat2`, and `dat3`. Each dataset corresponds to a separate exam question. (Thus, this exam has three questions.) You should find the best SARIMA$(p,d,q)\times(P,D,Q)_s$ fit(s) for each dataset (or a transformation thereof) that you can. *Your output should be in the following format. (Points will be deducted if it is not.)*

- The analysis for each question/dataset should begin on a new page and should have as label the name of the dataset ("dat1", "dat2", or "dat3").
- On the first page you should state on which page each question begins. [Note: One simple way to automate this in LaTex is to use `\section{}` to start each dataset and then include a `\tableofcontents`. You could also use `\label{}` and `\pagref{}` commands.]
- On the first page of output for each problem, you should first have a summary (labeled "Summary") that provides the model chosen, whether any transformation was used, parameter estimates, standard errors, and p-values in that model. Specify explicitly if you exclude a constant term. For example, "For the series $Y_t = X_t^{1/2}$, I chose an SARIMA$(1,2,3)\times(4,5,6)_7$ model, including the intercept term. The parameter estimates were ...". If you believe the data cannot distinguish between two

(or more) models you should describe both (all) of them in this manner here. (But if the grader(s) disagree then you may lose points.)

- After the summary, should be an explanation (labeled "Explanation"). Provide a clear explanation of why you selected the model you selected. Refer to the output of your analysis, which will be below. The model selection and diagnostic techniques we have discussed in class can be discussed here. You do not need to (and *should not!*) provide an exhaustive list of all possible models, but should rather provide explanation for which models were reasonable contenders (and why), and which model (or models) were the best out of those contenders (and why).
- After the explanation is the "Output" you refer to in your summary. (The output may be plots or output from various commands.) All of it should be clearly labeled or described. You do not need to provide exhaustive output from every command you have run, but you should include enough to justify all the arguments you make in your summary. (Please make extensive use of the `invisible()` command for capturing unnecessary output.)

Finally, *please refer to the original/raw (untransformed) time series as X_t in your descriptions and as xx in your code. Refer to any transformed series as Y_t in your descriptions and yy in your code*.

Questions.

1. Analyze Dat1.
2. Analyze Dat2.
3. Analyze Dat3.