Exam Instructions

Assigned: Thurs Dec 12, 2024 (3pm)
Due: Uploaded to Canvas on Tues, Dec 17, 2022, before 11:30am (*morning*, Central Time). Upload your exam as a single pdf file.
Late work: The Canvas final assignment is set to allow uploads up to a day after the actual due date/time (stated above). If an assignment is turned in late, 15 points will be deducted. No assignments more than 24 hours late will be accepted (and so you may fail the course if the assignment is not turned in on time and you have not communicated with me at all in advance!).

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 may not yield full credit.)

Note: statistics students have been caught cheating by working on takehome exams together and sadly have been forced to leave our program and university. 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.

General formatting guidelines:

The usual formatting rules:

- Your exam should be formatted to be easily readable by the grader.
- You may use knitr or Sweave in general to produce the code portions of the HW. However, the output from knitr/Sweave that you include should *be only what is necessary to answer the question*, rather than just any automatic output that R produces. (You may thus need to avoid using default R functions if they output too much unnecessary material, and/or should make use of `invisible()` or `capture.output()`.)
    - For example: for output from regression, the main things we would want to see are the estimates for each coefficient (with appropriate labels of course) together with the computed OLS/linear regression standard errors and p-values.
- Code snippets that directly answer the questions can be included in your main homework document; ideally these should be preceded by comments or text at least explaining what question they are answering. Extra code can be placed in an appendix.
- All plots produced in R should have appropriate labels on the axes as well as titles. Any plot should have explanation of what is being plotted given clearly in the accompanying text.
- Plots and figures should be *appropriately sized,* meaning they should not be too large, so that the page length is not too long. (The arguments fig.height and fig.width to knitr chunks can achieve this.)

Questions

On this takehome exam, you will analyze three data sets. Find `final-data.rsav` on the course webpage: there will be a link to it in the "Final" assignment on Canvas. Load `final-data.rsav` into R by running `load("path/final-data.rsav")` where "path/final-data.rsav" is replaced by the full path on your hard drive to the file final-data.rsav (the syntax for which is operating system dependent). The exam has three questions and the file has four data objects: : `dat1`, `dat2a` and `dat2b`, and `dat3`. Each dataset corresponds to a separate exam question. (Thus, this exam has three questions, with two parts for question 2.) Your output should be in the following format. (Points will be deducted if it is not.)

- The analysis for each question should begin on a new page and should have as label "Question 1", "Question 2", or "Question 3".
- On the first page of the exam, you should state on which page each question begins. [Note: One simple way to automate this in LaTeXis 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 a succinct answer to the question.

E.g., for SARIMA modeling, provide 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 ARIMA$(1, 2, 3)$ 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.

- After the summary, should be an explanation (labeled "Explanation"). Here you should provide a clear explanation of the methodology or reasoning that lead to your answer in the Summary above. Refer to the output of your analysis, which will be below. This part is still succinct (and does not include code output) but explains in words why you got the answer(s) you got.

  For SARIMA modeling, 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/will 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 single command you have run, but you should include enough to justify all the arguments you make in your summary.

Finally, *in Questions 1 and 2 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 Z_t in your descriptions and zz in your code. In Question 3, the two series are named xx and yy and you should not rename them.*

1 Question 1

(33 points)
For `dat1`, your job is to fit the best SARIMA$(p, d, q) \times (P, D, Q)_s$ model you can to the dataset.

Question 2

(33 points)
Consider `dat2a` and `dat2b`. There are two parts to this question.

1. For `dat2a`. Pretend we have a "baseline" model, which is AR(1) with $\phi_1 = .7$ and $\sigma^2 = 1$. Your job is to assess whether the true spectral density of the data match the spectrum from this AR(1) model or not, specifically at the following frequencies: $\omega = 0.027, 0.049, 0.25.$. That is, using (two-sided) 95% confidence intervals (CI's), test the hypotheses that the true $f(\omega)$ equals $f_{AR(1)}(\omega)$ (with $\phi_1 = .7, \sigma^2 = 1$) at the 5 omega values given. (Do not adjust for multiple comparisons.) [Note: to construct a hypothesis test from a CI, you use the natural 'duality' of CI's and hypothesis tests: you reject the null if-and-only-if the CI does not contain the hypothesized value.]

2. For `dat2b`. This data is identical to the previous one, except it has just one more observation. Now just answer the previous question again except for simplicity do this just for $\omega = .027$.

To format your answers, follow the same general formatting used for our analysis of SARIMA models: a Summary (describe just what your result is for each test), Explanation (the broad overview of the methodology that led you to your summary results), and Output. You may create one "Summary", "Explanation", "Output" section, and answer the two parts of this question in each, or you may have two separate "Summary", "Explanation", "Output" sections, one for each part. (Note: you may use any R functions you wish for this question, you do not need to do anything 'by hand'.)

Question 3

(34 points)
dat3 has two columns, yy and xx, which are both time series. Your job is to regress yy on xx using the ("transfer modelling") methodology we discussed in class.