STAT 520 – Fall 2023 – Test 2

Note: For this midterm exam, **you are not allowed to receive help from *anyone except me* on the exams.** For example, you may not talk to other students about the exam problems, and you may not look at other students' exams. **Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity.**

In addition, the writing of the mini-reports on this exam must be done entirely by you --- not with the help of any other individual or any AI program such as ChatGPT. You are welcome to use the textbook, course website, and other STAT 520 materials as aids in doing the problems. If you use other background sources (I'm not saying that this is necessary to do so), then please cite the sources you used.

Problems 1-3 below involve using R to analyze some real time series. You can run the R code given at the following web site to input the data and turn each vector into a time series object:
`http://people.stat.sc.edu/hitchcock/STAT520DataEntryRcodeTest2Fall2023.txt`

Your mini-reports should be typed in paragraph form and should include relevant graphs where necessary. While you can and should include graphs to supplement your analysis, **please do not clutter up your reports with R code and unedited R output**. For each report, the amount of actual text (not counting plots and graphs) does not need to be more than about one page in length.

Note that there may be several ways to satisfactorily answer these questions. In addition, since these are real data, it is possible that no model may *perfectly* describe the time series behavior. Your reports will be graded partly on the quality of the statistical analysis that you do, and partly on your ability to communicate your conclusions clearly and concisely. Specifically, each problem will be worth 20 points, for a total of 60 points:

Writing (out of 10 points): How organized, clearly written, comprehensible, and grammatically correct is the report? Would the client reading this report be confident that it was written by an educated, well-trained statistical scientist?

Analysis (out of 10 points): Were the graphs and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your statistical conclusions about the data set sensible and clearly justified by numerical or graphical evidence?

For each of the following data sets, you will conduct a complete analysis, including model specification, parameter estimation, model checking/diagnostics, assessing model fit, and relevant forecasting. Note that for some data sets, more than one model might be reasonable, so how you provide evidence to justify your choice of model is as important as which specific model you choose. You should consider aspects such as whether the time series process is stationary, and if not, whether it can be made stationary by some procedure, such as differencing. Also consider whether a transformation of the response variable is needed. It is recommended that for each analysis, you write the model equation of the model you chose, with estimated parameters plugged in.

1. Baseball playing styles and equipment have changed over time, but how have pitchers' performances changed, if at all? In this problem, we will analyze the major league baseball pitching performance over time. The data object `WHIP.ts` contains the WHIP values for major league baseball (all teams combined) for each year between 1871 and 2013. [WHIP stands for "Walks and Hits per Inning Pitched" and is a measure of pitching performance. The lower the number, the better the pitching performance.] Conduct and summarize a full analysis of the data. Augment your report with relevant graphics or plots, and be sure to comment clearly about what the graphs tell us.

Use your chosen model to obtain forecasts and 95% prediction intervals for the forecasted values for the next 10 years: 2014, 2015, …, 2023. About how many of these 10 prediction intervals would you expect to contain the true WHIP value for the corresponding year? Note that in fact, the WHIP values for the major leagues for years 2014, 2015, …, 2023 are: 1.275, 1.294, 1.325, 1.342, 1.304, 1.334, 1.327, 1.297, 1.266, 1.315. For your model and your prediction intervals, how many intervals contained the true value for that year? NOTE: Do not use these forecasts to calibrate your choice of model; the model selection should be done strictly based on the 1871-2013 data.
[Data from baseball-reference.com]

2. The Canadian lynx: This noble creature roamed the wild Northern tundra for centuries before being hunted by pelt-seekers in the 19th and 20th centuries. The object `lynx.ts` gives the annual number of lynx trapped in the McKenzie river district of northwest Canada, between 1821–1934. Conduct and summarize a full analysis of the data. Augment your report with relevant graphics or plots, and be sure to comment clearly about what the graphs tell us. Use your chosen model to obtain forecasts and 95% prediction intervals for the forecasted values for the next 3 years, specifically 1935, 1936, 1937.

3. It is said that home ownership is the American dream, but achieving that dream can be very expensive. The object `hp.ts` contains the annual median home price (in dollars, not adjusted for inflation) in the United States from 1963 to 2013. Conduct and summarize a full analysis of the data. Augment your report with relevant graphics or plots, and be sure to comment clearly about what the graphs tell us. Use your chosen model to obtain forecasts and 95% prediction intervals for the forecasted median home values for the next 9 years: 2014, 2015, …, 2022. A hypothetical question you should answer: Note that the last observed median home price, in 2013, rose above $300,000. Based on the observed data and your model, find the approximate predicted probability that the median home price in 2014 will be *more than* $350,000.

Note that the actual annual median home prices for 2014, 2015, …, 2022 were: 345450, 350450, 359650, 381150, 382475, 379875, 389800, 457375, 535500. Are your forecasts for the years between 2014 and 2022 that are based on the data through 2013 accurate or not? Make some comment about the accuracy (or lack of accuracy) of the forecasts over these years.

The midterm exam will be due by Friday, November 17 by 11:59 p.m.