2020 Applied Statistics Qualifying Examination

Instructions

- The exam commences on Tuesday, May 12 at 10am. The completed exam must be submitted as an email attachment to Jean McKee at `statsphdprogram@umich.edu` by 10am on Friday, May 15.
- During the examination period, you must not communicate with anyone other than Kerby Shedden, Gongjun Xu, or Ji Zhu about the exam or about any topics or methods related to the exam.
- You are free to use any media resource during the exam including course notes, books, and web sites. You are free to use any computing resources or software tools.
  You **must** cite all sources that you use, just as you would when writing a paper.

The data for the exam are mortality counts for the United States from 2007 to 2018, provided by the US CDC (Centers for Disease Control). The original source for the data is at the link below.

`https://www.cdc.gov/nchs/nvss/mortality_public_use_data.htm`

For this exam you should work with a consolidated version of these data available here:

`https://umich.box.com/s/mxo8sdjsf0r5mwae1n858zqbb29pd1vb`

The data description that we provide below should be sufficient for completing this exam, however if you are interested, the complete data documentation from the CDC is available at this link:

`https://www.cdc.gov/nchs/nvss/mortality_public_use_data.htm`

The *Deaths* variable is a count of all deaths in the US of resident persons, in a given demographic cell, in a given month of a given year. The demographic variables are age (binned into 18 5-year intervals) and sex (female or male). For example, the first row of the dataset records 1087 deaths for females under age 5, in January of 2007. Months are coded "1" for January, "2" for February, etc. The *Population* variable reports the total US population for the given age × sex demographic cell in the given year. Note that the population is not specific to the month.

You are free to pursue any meaningful question that can be addressed using these data. A few fruitful directions to consider, alone or in combination, are listed below. These are not intended to be exclusive and you are welcome to pursue other directions.

- What are the demographic and/or temporal factors associated with higher versus lower mortality?

- What is the seasonal pattern of mortality? Is it sex-specific and/or age-specific? Is it stable over time?
- What trends in mortality, if any, are seen in the 12 years covered by these data?
- What patterns of independence or conditional independence are reflected in the data?
- After accounting for major systematic trends, do the data follow a Poisson distribution?
- What patterns of serial dependence, if any, exist in the data for one demographic cell viewed as a series of observations indexed by time?
- Are there any exceptional values that do not fit the patterns implied by the majority of the data?

There are many promising possibilities for analyzing these data using the methods discussed in Statistics 600 and 601. We ask that you primarily make use of the methods emphasized in those two courses.

Expectations

- You should prepare a written report that is 10 – 12 pages in length including all tables and figures and with 1.5 – 2 line spacing. The report should be submitted as a single PDF file.
- Do not include any computer code in your main document. You may also provide an appendix with additional output and/or code but we are not obligated to consider the appendix when evaluating your exam. If you choose to provide an appendix please submit it as a separate PDF file.
- This is primarily a data analysis exam. You should focus on a limited number of questions that you think an applied researcher would ask based on these data. Motivation for the questions that you address should be clearly stated.
- The questions that you pose should primarily be addressed using statistical methods covered in **Stats 600 and 601**. You are strongly discouraged from using more advanced methods, or novel methods that you have invented for this project.
- You should motivate any data-analytic techniques that you use, and describe any limitations of your analysis. You should also discuss the implications of your results.
- A strong project does not necessarily need to identify a strong predictive relationship or a statistically significant association. An interesting question may turn out to have an ambiguous answer given these data. In that case, it is important to explain what limitations of the data or methods may have led to the outcome being ambiguous.
- The main review criteria are:
  1. Your report is clearly written.

2. Your report addresses a clearly-stated and focused research question.
3. Your report reflects a deep understanding of some of the techniques covered in the Stats 600/601 courses.
4. All of the claims made in your report are supported by the data, through your analyses.