Qualifying Exam 2017

Take Home Question 1

This part of the exam requires you to describe, execute, and interpret a data analysis. These data come from the Cebu Longitudinal Health and Nutrition Survey. The CLHNS was conducted by a team of researchers from the United States and the Philippines, the CLHNS is part of an ongoing study of a cohort of Filipino women who gave birth between May 1, 1983, and April 30, 1984. The CLHNS was originally conceptualized as a study of infant feeding patterns, particularly the overall sequencing of feeding events (i.e., of both milk and non-milk items), the various factors affecting feeding decisions at each point in time, and how different feeding patterns affect the infant, mother, and household. The idea was to study these topics within as natural a setting as possible and to analyze how infant feeding decisions by the household interact with various social, economic, and environmental factors to affect health, nutritional, demographic, and economic outcomes. The cohort of children born during that period, their mothers, other caretakers, and selected siblings have been followed through subsequent surveys conducted in 1991-92, 1994, 1999, 2002, and 2005. Research is now focused on the long-term effects of prenatal and early childhood nutrition and health on later adult outcomes, including education and work outcomes and development of chronic disease risk factors. Three Data Files are included in the github repository:

- `data_2002.rda`
- `data_2005.rda`
- `data_2007.rda`

We will focus on self-rated health, which is a metric of how an individual perceives their health status, typically on a scale from poor to excellent. As just a single question, self-rated health (SRH) is easy to collect and is widely used on health questionnaires. Despite its simplicity, it provides valuable insights to health researchers and practitioners. SRH allows individuals to incorporate their personal judgments and attitudes towards health, and thus, SRH can capture health in a more holistic manner than objective health measures. Your goal is to model SRH in the 2002, 2005 and 2007 CLHNS data. This data is longitudinal in nature and there will be many ways to analyze this. Follow these steps as you model this data:

1. Explore the data in all 3 years.
2. Model SRH using the 2002 data as training data.
3. Predict SRH in 2005 and 2007.

While there are many different ways you can model the data, you will not have time to try everything. Develop a reasonable model and if there are additional forms you do not have time to try but believe would be important to investigate, describe them in your report as future work that could be done. Be sure to discuss the accuracy of the model built using the 2002 data when you predict into 2005 and 2007.

Your responses should be self contained so that readers do not have to search through the exam to find relevant output. Feel free to cut and paste from computer output, but please make sure tables and/or graphs are properly formatted and easy to read. Be succinct. Your report probably does not need to exceed 5 pages. You do not need to summarize every step you took, but should describe clearly your final results including:

- Description and justification of analytic method
- Summary of key findings with a plain language summary abstract that explains major results to a non-technical reader.
- Be sure to
- Tables and/or figures as needed
- An appendix that includes the computer programs used to carry out the analyses. Please annotate the program to indicate which parts were used to answer specific items. This will help the grader when an answer differs from what is expected.

Files and Attachments

1. Data Folders

- The folders marked for the individual years contains the tab separated files of the different databases used in the data.

2. Data Scripts

- These files display the pre-processing and cleaning which has already been completed on the data files that you will be using.
- Consult this in order to see what variables are included in the data.
- The particular `variables_200*.txt` files will display what variables were placed into the data files.

3. Codebook
    - There is a folder labeled codebook. This contains the codebook for the particular combined questionnaires from 2002, 2005 and 2007.
4. `data_200*.rda`
    - These are the individual data files for each of the given years.