PSTAT 220B, Winter 2020, Final Project

Due uploaded to Gauchospace before 4 pm on Thursday, March 19, 2020.
Early projects are welcome!

General Instructions

Important Final Project Rules: This project is treated as an exam, therefore all data analysis and report writing must be your own. No one except Professor Meiring or Caroline Yovanovich may provide any guidance on your project, including writing. You may not discuss any part of the data analysis with others, and no one is allowed to correct your writing. Evidence of collaboration with other prior or current students, or with other individuals, will be severely penalized. However, you may use any reference books and your own notes from PSTAT 220B and other courses.

Your answer should be in the form of a well-structured report up to maximum length of 12 pages including all figures and tables. This main report must be uploaded in pdf format, although additional appendices may be .Rmd or similar files. Your report should be written to a pharmacy manager who has some basic familiarity with statistics (at the level of PSTAT 5LS or PSTAT 5A), but who has not taken advanced statistics courses. Remember to explain all notation and formulae, which methods you used and why; results and interpretations; conclusions. Tables and graphs should be clearly labeled, with clear informative captions. Answers written on computer code are not permitted as part of your main report, however in an additional appendix, you may include *annotated computer results*, but do not expect that this appendix will be read. Your main report (up to 12 pages) must be understandable when read without this additional appendix.

It is very important to remember that the reader will neither be familiar with R model statements, nor the notation that R uses in computer results. Therefore don't simply paste R output into your report, since this will include notation that is not familiar to the reader. You must write all models and assumptions in clear equations and clearly define your notation so that your models and assumptions (and any parameter estimates, figures, table headings etc) will be understandable to the reader. Mathematical notation is fine in your report, provided it is explained.

With best wishes for your data analysis and report writing!

For this project, suppose a survey was conducted within San Francisco to investigate the number of individual consultations with pharmacists. The following variables were collected for 500 surveyed individuals, and provided to you in the dataset *pharmacist.txt* posted on Gauchospace.

pc	Number of consultations with a pharmacist in the past 4 weeks
sex	1 if female, 0 if male
age	Age in years divided by 100
income	Annual income in US dollars divided by 100000
lp	1 if covered by private health insurance with pharmacy coverage, 0 otherwise
fp	1 if covered by private health insurance without pharmacy coverage, 0 otherwise
fr	1 if not covered by private health insurance, 0 otherwise
ill	Number of illnesses in the past 4 weeks (5 or more illnesses are coded as 5 in this dataset)
ad	Number of self-reported days of reduced activity in the past 4 weeks due to illness or injury
hs	General health questionnaire score calculated by the investigators. A high score indicates poor health.
ch1	1 if individual has chronic medical condition(s) but is not limited in activity, 0 otherwise
ch2	1 if individual has chronic medical condition(s) that limit individual's activity, 0 otherwise

Analyze these data and write a report to

- Investigate how the number of consultations with a pharmacist is associated with other variables. As part of this investigation, develop a model for pharmacy managers to estimate the expected number of pharmacist consultations by a new customer within a 4 week period, if the pharmacy was provided with values of the other variables for that new customer.

- For the first person in this dataset, use your fitted model to estimate the probability distribution for his/her number of pharmacy consultations within a 4 week period, i.e., estimate the probability that his/her number of pharmacist consultations equals 0,1,2, etc.

- Clearly state and discuss your model(s) and assumptions, and also the limitations of your analysis in the context of this study. Within the discussion section, you may include any questions you would like to ask the people who designed the survey and collected these data.

When writing your report, aim your report at a pharmacy manager who has some basic familiarity with statistics (at the level of PSTAT 5LS or PSTAT 5A), but who has not taken advanced statistics courses. Note: I only expect you to use GLM and related methods studied in the course before we started the log-linear models set of notes for categorical data, in addition to exploratory data analysis, and any methods from pre-requisite classes that you also find helpful. Refer to the previous page for additional instructions on the report requirements.

Note: these data/descriptions have been modified for this class final project, and do not represent an actual survey run in San Francisco. Not to be used beyond this course.