STAT 796: Final Exam

This exam is in the form of a take-home data analysis project, with two components. Each part of the assignment addresses a different analytical goal: (1) Evaluating Risk Factors for Heart Disease and (2) Predicting Heart Disease.

Background: Heart disease comprises a major morbidity and mortality burden in the United States, accounting for over 600,000 deaths every year–almost 1 in 4 deaths[1]. Coronary heart disease (CHD), the most common type of heart disease, occurs when blood vessels become partially or fully obstructed, reducing the flow of blood to the heart and other areas of the body. CHD often leads to myocardial infarctions (MI), which are commonly called heart attacks. MIs occur when the heart receives insufficient oxygen and can lead to major damage to the heart and to death. Symptoms of CHD include angina (chest pain) and shortness of breath, although CHD often goes undiagnosed until an MI occurs. CHD can also lead to congestive heart failure (CHF), which occurs when the heart muscles (ventricles) cannot contract or relax enough to pump blood like normal. If risk factors for CHD can be identified, then these risk factors can be addressed in individual treatment plans and in public health policies to reduce the incidence of CHD.

The Framingham Heart Study (FHS) is a landmark long-term prospective study of cardiovascular disease among individuals living in Framingham, MA. Enrollment began in 1948 with 5,209 subjects initially enrolled. Participant data has been collected biennially since enrollment, with extensive adjudication of cardiovascular outcomes from sources including participant contact, death certificates, and hospital records. The original FHS has been expanded multiple times since study inception to include more diverse groups of individuals, but for the present purposes we consider the original cohort.

Part 1: Evaluating Risk Factors for Heart Disease

Analytical Goal: Using data from FHS, design and conduct an analysis to answer the following questions:

1. Are serum cholesterol levels associated with risk of MI or fatal CHD?
2. Is there evidence of an association between cholesterol levels and time until MI or fatal CHD?
3. Are there other risk factors for MI and fatal CHD?

Response Format: Your answer to these questions should be constructed as a report that contains the following components:

- **Abstract**: A one paragraph summary that briefly describes the questions of interest and summarizes your overall conclusions and results, as one might see in the abstract of a published article in an epidemiology journal.
- **Introduction**: A short (max 2 paragraph) explanation of the background and context for the questions of interest. For the purposes of this assignment, you do not need to conduct a literature search or otherwise bring in additional information. You may use the information from this prompt for the introduction, but translate into your own words and understanding.
- **Variable Selection**: A discussion of the relevant variables for addressing the questions of interest, including which version(s) of the cohort data you will use. For each variable in the dataset, address its relevance or lack thereof, including potential confounding. (*You are not expected to have the expert knowledge of a cardiologist and while citing literature or external sources is permissible, it is not required. You are expected to present what you believe to be plausible relationships based upon your own knowledge and the examples we have seen in class.*)
- **Statistical Methods**: Describe the statistical method(s) you will use to answer the questions of interest. Specify the model(s) that will be fit, the form of variables included in the model(s), and what

[1] https://www.cdc.gov/heartdisease/facts.htm

hypothesis test(s) will be conducted to answer the questions of interest. Distinguish between analyses that are pre-specified and those that are not.

- **Descriptive Statistics**: Provide numerical and/or graphical summaries of the cohort, including relationships relevant to the questions of interest. For example, this could include summary statistics stratified by the outcome variable. Provide a prose description of these summaries, as one might see in an epidemiological journal article.
- **Results**: Present and explain the quantitative results of your statistical analyses, as one might see in an epidemiological or statistical journal article. If you provide an estimate, explain what the quantity represents and provide a measure of uncertainty. If you provide the result of a statistical test, explain what was being tested.
- **Discussion**: Provide a qualitative summary of your results and answer to the questions of interest. Discuss any limitations of your reported results, recommendations and insights from sensitivity analyses, or other matters relevant to the questions of interest.
- **Appendix**: An optional appendix, included a separate section at the end of your report, that contains figures or tables that are discussed elsewhere in the report. Placing figures/tables in the appendix makes the main report easier to read, although the main report can include a small number of figures and tables inline. For example, if have many figures to discuss from a sensitivity analysis, place them in the appendix and reference them in the main text earlier.
- **R script**: A *separate* .R file containing R code corresponding to the report. This should include executable code for producing the summary measures and model fitting. Labels sections of the code with comments so that it is readable.

Note:

- Questions 1 and 2 ask for a specific confirmatory analysis, while question 3 is more exploratory.
- Only one report is needed for addressing all three questions, although you may wish to have separate methods and results sections for the different questions of interest.

Part 2: Predicting Binary Outcomes

Analytical Goal: The researchers would also like to develop a model for predicting a MI or fatal CHD in patients. This model will be used to estimate risk for individual patients based upon their measured or reported characteristics. Using data from FHS, develop a logistic regression model for predicting whether or not a patient will have an MI or fatal CHD within the next 20 years.

Response Format: Your response to this prompt should be constructed as a report that contains the following elements:

- **Abstract**: A one paragraph summary that briefly describes the analytical goal and summarizes your overall conclusions and results.
- **Model Selection**: Describe the procedure you use to select a model. This should include both the overall procedure you use and what measure(s) of performance are used for the selection.
- **Model Performance**: A graphical and/or numerical description of the performance for your chosen prediction model. Evaluation of the performance of your chosen model should includes its predictive accuracy for observations in the dataset `fh_20yr_test.csv`
- **Cutpoint Selection**: Choose a reasonable cutpoint for your prediction model, and explain your choice. Describe how well the model predicts when using this cutpoint and any limitations of this cutpoint.
- **R script**: A *separate .R file* containing R code corresponding to your model development and evaluation. Labels sections of the code with comments so that it is readable.

Additional Information

Dates and Deadlines: The assignment description and data will be posted on Canvas, and the reports should be submitted on Canvas as PDF or Word files (with separate ,R script file for each part).

- Exam prompts released: Wednesday, April 24
- Data for Parts 1 & 2 released: Monday, April 29
- Test Data for Part 2 released: Tuesday, May 7
- Final Reports Due: **Friday, May 17 at 12:00pm**

Data:
The available variables for the FHS cohort are provided in the following table. All values are from baseline, expect for those indicating presence or time of MI and fatal CHD events.

Variable	Description	Units and Coding
`id`	Identification number for each participant	NA
`sex`	Sex of participant	0 = male, 1 = female
`age`	Age	Years
`sbp`	Systolic Blood Pressure	mmHg
`dbp`	Diastolic Blood Pressure	mmHg
`bp_meds`	Taking Blood Pressure Medication	0 = No, 1 = Yes
`smoke_curr`	Current Smoker	0 = No, 1= Yes
`cigs_day`	Cigarettes Smoked per Day	Count

`educ`	Highest Education Level	`Less than High School`
		`High School / GED`
		`Some College`
		`Bachelors or higher`

`chol`	Serum cholesterol level	mg/dL
`bmi`	Body Mass Index	kg/m^2
`gluc`	Serum glucose level	mg/dL
`diab`	Diabetes Status	0 = No, 1 = Yes
`hr`	Heart Rate	Beats/min
`prev_angina`	Existing Angina	0 = No, 1 = Yes
`prev_ht`	Existing Hypertension (High blood pressure)	0 = No, 1 = Yes
`prev_stroke`	Prior Stroke	0 = No, 1 = Yes
`mi_fchd_in20`	MI or Fatal CHD within 20 years of baseline	0 = No, 1 = Yes
`mi_fchd`	MI or Fatal CHD	0 = No event, 1 = Event
`time_mifchd`	Time of MI or Fatal CHD event, or time of censoring	Days

Three datasets are available for you to complete this project:

- `fh_20yr.csv` – Data on subjects who had at least 20 years of follow-up or had an MI or fatal CHD within 20 years of baseline. Use this for Parts 1 and 2.
- `fh_mifchd.csv` – Data on subjects with any amount of follow-up. Use this for Part 1.
- `fh_20yr_test.csv` – Data on 500 subjects that were randomly excluded from the other two datasets. Use this for out of sample validation for your prediction model in Part 2.

Note: Although the data for this assignment have been modified from the original records to protect confidentiality, these data should not be used for any purpose, current or future, other than the completion of this assignment.

Evaluation: This is an individual exam and you are expected to work independently. While you may consult course notes and other supporting materials, do not use any published references from the literature that contain results from the FHS.

The total point value for this project is 70 points for Part 1 and 30 points for Part 2, for a total point value of 100 points. Evaluation of your report will be based on the following factors:

- Completeness of response to the specified analytical goals
- Completeness of all required sections in both reports
- Appropriateness of the chosen statistical methods
- Choice and explanation of model selection procedures
- Explanation of rationale for confounding adjustment (Part 1)
- Choice and explanation of model evaluation procedure (Part 2)
- Interpretation of results, including point estimates, measures of uncertainty, and statistical tests
- Overall writing (be clear, concise, and grammatically correct)
- Appropriate labels, captions, and units for all tables and figures