The University of Minnesota
School of Statistics

Ph.D. Written Examination
Applied Exam

From (9 am) Tuesday, August 25, 2020
To (11:59 am) Thursday, August 27, 2020

This part of the exam consists of two data analysis problems. Please submit your solution electronically, as two .pdf files, to Taryn Verley (`tverley@umn.edu`) by 11:59 am on Thursday, August 27, 2020.

Rules:

- Please submit a separate .pdf file for each of the two problems.
- Please put your 3-letter code name (instead of your real name) into the solution preamble AND filenames.
- The applied exam is open book, open notes. You can use the internet. However, all work needs to be your own. Academic misconduct will result in failing this examination.
- Do not consult with any person, except the faculty member proctoring the problem, or questions about Problem 1 to Professor Birgit Grund (`birgit@umn.edu`) and questions about Problem 2 to Professor Lan Liu (`liux3771@umn.edu`).

Problem 1

Blood banks harvest and freeze peripheral blood stem cells for leukemia patients for "autologous donations", meaning that the patient donates his or her own cells, and he or she will be the only one who will receive the cells in the future. More accurately, blood banks harvest a large number of cells from the peripheral blood, with the goal to have a sufficient number of peripheral stem cells, which tend to be CD34+ cells. When the patient dies, any remaining blood products for the patient are destroyed.

These CD34+ collections have been going on for decades, but in 2017, the accreditation agency for the blood bank required that the blood banks perform "stability studies" of the saved products. **The goal of the stability studies was to investigate for how long the blood products retained their effectiveness, and establish an expiration date when effectiveness is lost.**

The data for this problem are similar to a sample data set from a blood bank, and you are asked to perform the stability study. Data are available for 37 patients, enrolled between 2008 and 2019. Prior to 2017, the blood bank would extract two aliquots (essentially small subsamples) of the CD34+ product at the time of collection. They would "count" one aliquot at the time of collection, and they would "count" the second aliquot at the time a product was destroyed. Beginning in 2017, the blood bank would extract six aliquots at the time of collection. These aliquots would be "counted" at time of collection; six months, one year, three years, and five years after collection; and at time of destruction. By "counting", the following is determined:

1. The percentage of cells (leukocytes) that are CD34+ cells.
2. The percentage of CD34+ cells that are viable (alive).

At the time of collection, the vast majority of cells are viable (usually over 97%), and the percentage CD34+ is usually between 0.5% and 1%, depending on the individual. Theoretically, it would be expected that cells are loosing their viability over time. **Assume that the blood product has lost its effectiveness if the percentage of cells that are viable CD34+ cells is below 0.05%.**

The data have the following quirks:

- There was an instrument change on December 31, 2016, which may affect the way how CD34+ cells are counted, and how "viability" is measured.
- Most patients have given only one blood donation ("sample"), but some have given two donations.
- The number of viable cells in the product should decrease over time, because cells decay. However, the number of cells that are "counted" may decrease or increase over time; there are biological reasons, but let's consider an increase as instrument failure – it is not real.

The data set for this problem is in the file `CD34.rda`, available here. The data file is also attached in the exam package. It has the following variables:

Name	Values	Description
Patient	integer	Patient identifier
Sample	1, 2	Identifier of sample within Patient
Coll.year	2008-2019	Year of blood collection
Stored	continuous	Time the blood product was stored at the time of measurement (years) 0.003=Day of collection, 0.5, 1, 2, ..., 10 years
Viability	continuous	Percent of CD34+ cells that are viable
CD34pct	continuous	Percent of cells that are CD34+ cells
viableCD4pct	continuous	Percent of cells that are viable CD34+ cells (**CD34pct*Viability/100**)

The ultimate goal of your analysis is to perform the stability study based on the available data.

(a) Describe the effect of the instrument change on the measurement of the *percentage of cells that are viable CD34+ cells*. Estimate a conversion equation that converts **viableCD34pct** values measured with the old instrument (prior to 2017) to values that would have been measured with the new instrument. Quantify the uncertainty in your conversion (e.g., with a 95% CIs). Describe the assumptions that you used to derive the equation; are the assumptions fulfilled?

To help guide you, below are a few items that you should include in your report.

- Are there outliers? If yes, justify how you treat them in your analysis.
- Blood samples where the entire series was measured with the new instrument (blood donated in 2017 or later) were stored for at most 3 years, while blood samples where the series was measured with the old instrument may have substantially longer storage times. Can you reasonably combine measures done with the old instrument and measures done with new instrument to estimate the trajectory over time and the instrument effect? If yes, how? If no, why not?
- Justify your model choice.

(b) Describe the trajectory of decay in the *percentage of cells that are viable CD34+ cells* over time.

(c) For a blood product collected in 2019, predict for how long the product will retain its effectiveness (number of years). Quantify the uncertainty that the blood product is still effective after 1 year. A blood product is effective if the *percentage of cells that are viable CD34+ cells* is >0.05%.

Present your main results and justifications in a way so that non-statistician colleagues could understand. **Summarize your results in at most 5 pages, plus at most 5 pages of supporting tables and figures.**

Hint: Do the best you can with the data you have. This is a real-life problem and the data is not ideal. The goal is to provide something useful and reasonably correct for the people at the blood bank.

Problem 2

The American Heart Association provides three categories for total serum cholesterol levels:

1. Less than 200 mg/dL: "Desirable level" – an individual at this level is at lower risk for coronary heart disease.

2. 200 to 239 mg/dL: " Borderline high" – an individual with a cholesterol level of 200 mg/dL or higher has increased risk.

3. 240 mg/dL and above: "High blood cholesterol" – an individual with this level has more than twice the risk of coronary heart disease as someone whose cholesterol is below 200 mg/dL.

The investigator is primarily interested in understanding the relationship between an individual's body mass index (BMI) and his/her cholesterol levels, cross-sectionally and across time, with secondary interest in the relationships of age and gender with cholesterol level. The individuals were recruited from six health centers in Minnesota (A, B, C, D, E, F) and their ID begins with the corresponding center.

The dataset chol_baseline.RData contains data on 171 participants from an observational study. Each individuals also have their gender and age group recorded. The study is cross-sectional (we decided to collect 171 individuals and then cross-classified them).

For this exam question, you are being asked to examine how an individual's cholesterol level is affected by his/her BMI, gender, and age. "Undesirable" cholesterol levels are defined as being either "borderline high" or "high".

- ID: unique participant identification code. The first letter stands for their centers.
- CHOL_LEVEL: participant's cholesterol level (1=desirable; 2=borderline high; 3=high)
- BMI: participant's BMI for this visit
- AGE_GROUP: participant's age group at the study enrollment (1=55 and under; 2=over 55 but not exceeding 65; 3 = over 65)
- GENDER: participant's gender (M=male, F=female)

1. Assume that individuals in the same centers are independent. Answer the following questions.

   (a) Suppose the investigator is interested in the independence between having an undesirable level of cholesterol and gender, ignoring all other variables. Carry out a formal test. Clearly state the null and alternative hypotheses, the test statistic, the $p-$value, and a 95% confidence interval associated with the test statistic. If there is any additional assumptions, please state clearly. Summarize your conclusion to the investigator.

(b) The investigator is also interested in the independence between having an undesirable level of cholesterol level and gender among those in center B, ignoring all other variables. Carry out a formal test. Clearly state the null and alternative hypotheses, the test statistic, the $p-$value, and a 95% confidence interval associated with the test statistic. If there are any additional assumptions, please state clearly. Summarize your conclusion to the investigator.

2. Assume that individuals in the same center are independent. Fit a logistic model to whether participants have an undesirable cholesterol status, examing the effects of age, gender, and BMI, and also considering for inclusion of any interactions that you judge to be important. Write out your model clearly and justify your choice for the final model.

3. Assume that individuals in the same center are independent. Fit a multinomial logit model to cholesterol level (CHOL_LEVEL) using the "desirable" level as the reference level. Incoporate age, gender, and BMI as explanatory variables, and include any important interactions. Write out your model clearly and justify your choice of your final model.

4. The investigator finds the model in part 3 is of more interest. Provide one to two concise paragraphs suitable for the Methods section of a manuscript that clearly describes the model's fit to the data. In addition, provide two to three paragraphs (plus some related tables or figures) for the manuscript's Results section that give a comprehensive summary of your findings, including the interpretation of noteworthy odds ratio estimates (and the corresponding 95% confidence intervals); for significant interaction effects, if any, please also fully interpret the meaning of that interaction.

5. Suppose the investigator now tells you that individuals from the same center are given the same diet (details of the diets are not available) and the entire study was carried out in 200 centers across the USA. Explain how you would modify your model in part 2 to accommodate the dependency if you were given the study data from 200 centers. Please clearly write your models and assumptions if there are any.