Survival Analysis Statistics 255

Final Exam

Handed out: Wednesday, March 15, 2023 Due: Friday, March 24, 2023 at 10:00am

General Instructions: You may not communicate with anyone (in person, by phone, email, text or otherwise) about your exam other than me. You are free to look up any other references you wish as long as those references do not require you to communicate with another human being (besides me). If there are questions that you have about the exam, feel free to contact me either by email, phone, or in the office.

Turning in Your Exam: Exams will be turned in via Canvas. Your final report (contents described below) must be in a pdf format. I would also ask that you submit the R code used to generate the dataset and fit your primary models for Questions 1, 2 and 3. No late exams will be accepted and the resulting grade on a late final will be 0.

1 Background

End-stage renal disease (ESRD) is a condition where the filtration performed by the kidneys has been reduced to a point at which life can no longer adequately be sustained. According to data from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) over 850,000 persons in the United States are being treated for ESRD and many more suffer from early stage chronic kidney disease. The standard of care for adult ESRD patients that do not have access to a viable transplant is hemodialysis.

It has been postulated that indices of protein-energy malnutrition (PEM) may be markers of total mortality among hemodialysis patients. Serum albumin is a protein biomarker and surrogate for nutritional status that is among the indices of PEM that have been hypothesized to be associated with mortality. It is natural to hypothesize that serum albumin may be indicative of increased mortality because low albumin implies nutritional instability that may be due to inadequate dialyzing. However, albumin has also been associated with other risk factors for mortality including age, gender, smoking status, obesity, duration of renal disease, and cardiovascular disease. Further, because albumin levels tend to be affected by cardiovascular disease it has also been hypothesized that the association between albumin and the risk of mortality my differ between patients with and without cardiovascular disease. If serum albumin were established as an independent biomarker, this would provide nephrologists with an additional measure to monitor hemodialysis patients and potentially decrease the risk of mortality in these patients. Quantifying the association between albumin and mortality is the main objective of our analysis.

To address this hypothesis, $N = 1,979$ hemodialysis patients were recruited to an observational study. Patients who were undergoing dialysis during December of 2014 were eligible for the study and were recruited from dialysis clinics across the United States. At the time of recruitment, serum albumin was measured on all patients in addition to multiple demographic and laboratory measurements. Among these are age, race, gender, body mass index, smoking history, subjective assessment of nutritional status, history of cardiovascular disease, and laboratory measurements that have been associated with cardiovascular disease (the leading cause of death in dialysis patients). In addition, given the study's focus on albumin, albumin was also measured and recorded on patients every 1-2 months for up to one year following the start of the study.

2 Available Data

As noted above, data are available on N=1,979 patients undergoing dialysis therapy during the month of December 2014. Two data sets are available. The first dataset contains survival data and all baseline data (ie. measured at the time of recruitment into the study) and is located on the course website and is entitled `usrdsData.csv`. The second dataset contains followup albumin measurements that have been taken every 1-2 months for up to one year after the start of the study. Note that the first measurement of albumin in this dataset occurs at day 0 for each patient and represents the baseline albumin data given in the first file. The second dataset is located on the course website and is entitled `LongitudinalAlbumin.csv`.

A brief description of the variables in the file `usrdsData.csv` is as follows:

`usrds_id` :	unique patient identifier
`tdeath` :	end of observation time (in days; measured from time of recruitment into the study)
`death` :	indicator of whether the patient was truly observed to die
`age` :	age of the patient (in years) at the time of recruitment into the study
`female` :	indicator of whether or not the patient was female
`racegrp` :	race of the patient (1=caucasian, 2=African American, 3=other)
`smokegrp` :	smoking status of the patient measured at time of first access placement (1=never smoked, 2=former smoker, 3=current smoker)
`hist.cvd` :	indicator of whether patient has had a history of cardiovascular disease
`diabetes` :	indicator of whether the patient is diabetic (includes type I and type II)
`esrdtime` :	total time (in years) that patient had ESRD prior to entering study
`undnour` :	indicator of whether the patient appeared undernourished to the study nurse
`bmi` :	body mass index of the patient measured at time of recruitment (calculated as weight (kg) / height^2 (m^2))
`albumin.0` :	serum albumin measured in the patient at the time of recruitment (g/dL).
~~`cholest` :~~	total serum cholesterol measured in the patient at the time of recruitment (mg/dL). Total cholesterol has been shown to be associated with cardiovascular disease.
~~`trigly` :~~	serum triglycerides measured in the patient at the time of recruitment (mg/dL). Has been associated with cardiovascular disease, but evidence is weak conditional upon cholesterol and is generally highly correlated with serum cholesterol.
~~`pst.sbp` :~~	Post-dialysis systolic blood pressure measured at the time of recruitment (mmHg).

The file `LongitudinalAlbumin.csv` contains three variables: `usrds_id` (unique patient identifier that matches the baseline datafile), `albumin` (the longitudinally recorded albumin measurement), and `measday` (the study day that the albumin measurement was taken on).

The first ten lines of each of the the datasets are given by the following:

```
> usrdsData[1:10,]
   usrds.id tdeath death age female racegrp smokegrp hist.cvd diabetes
1      1179    427     0  69      1       1        1        1        1
2      1783    427     0  41      0       1        2        0        0
3      2015    427     0  76      1       2        2        1        1
4      2929    427     0  71      0       1        1        1        0
5      2953    169     1  62      0       1        3        0        1
6      3219    427     0  62      0       2        2        1        0
7      3499    427     0  57      0       3        1       NA        1
8      3525    427     0  73      1       1        1        1        0
9      3809    419     1  70      0       1        2        0        0
10     4138    427     0  69      1       2        1        1        0
```

```
   esrdtime undnour    bmi albumin.0 cholest trigly pst.sbp
1   2.84110       0 28.195       3.6     215    162     160
2   2.21370       0 20.988       4.1     102    195     160
3   0.90411       1 22.822       3.7     196    118     148
4  14.92603      NA 18.993       3.6     125     NA     134
5   2.86575       0 24.624       4.0     246    201     144
6   1.57534       0 30.664       4.4     162    277     166
7   6.81096       0 21.193       3.9     103     73     150
8  14.59726       1 16.230       3.0     145    116     140
9   4.41918       0 26.015       3.8     234    168     120
10  2.84658      NA 20.812       2.9     272    202     110
```

```
> LongitudinalAlbumin[1:10,]
   usrds.id albumin measday
1      1179     3.6       0
2      1179     3.9      28
3      1179     3.8      71
4      1179     3.9     134
5      1179     3.2     211
6      1179     3.3     260
7      1179     3.1     314
8      1783     4.1       0
9      1783     3.6      28
10     1783     4.0      84
```

3 Scientific Goals

Using the available data, you should employ appropriate regression modeling strategies (adjusting for covariates as you determine necessary) to address the following aims:

1. **Using only baseline data (ie. measured at the time of study start)**:
   (a) Quantify the association between baseline serum albumin and the risk of mortality.

(b) Quantify the potential difference in the association between baseline albumin and the risk of mortality comparing patients with and without a history of cardiovascular disease.

2. **Using data on repeated albumin measurements:** Repeat your analysis above to by considering albumin as a time-dependent covariate and compare/contrast your results commenting on the likely reasons for any differences you observe. (Hint: You may find the `merge()` function in R along with the `by=` option useful for this part of the problem.)

3. **Alternative sampling strategy:** Repeat your analysis for 1(a) using a nested case-control design with $M = 4$ controls per case. Compare your estimated coefficients to those obtained using the full cohort in 1(a) and comment on the savings in total of individuals used in the analysis relative to the increased variance observed for the primary question of interest. This section can be placed in the Appendix of your report.

4 Report

Your analysis results should come in a final report limited to 10 pages in length (excluding appendices if necessary). The format and structure of the report should follow the instructions at the end of this document. Grading for the exam will be based on the following criteria:

1. Scientific approach
   (a) Did you consider/investigate problems in the sampling that might materially affect the results?
   (b) In addressing each of the questions, did you choose appropriate models to answer the scientific questions?
2. Statistical approach
   (a) Were the methods chosen appropriate for the data at hand? Were any key assumptions violated?
   (b) Were the methods chosen reasonably efficient?
3. Written report
   (a) Were your findings well documented in a succinct manner?
   (b) Was the report written at an appropriate level?

Instructions for your Final Report

You are required to write-up the results of your analysis in a final report. In this report, you should describe the results of your analysis and the conclusions you would reach from those results. This report should look like a formal report to a statistically naive researcher or an interested lay person. Because a statistical analysis aims to answer a scientific question, you should organize your report in the manner which is customarily used in science. To wit:

1. **Abstract**: Provide a concise description of the question, the data used to try to answer it, and the conclusions of your analysis. Give the most pertinent estimates, confidence intervals, and P values. Don't give too much detail here, but do note any significant problems that were encountered.

2. **Background/Introduction**: Provide a description of the scientific motivation for the analysis. Use your own words rather than copying the description provided by me. By providing your understanding of the problem, I may be able to correct (take into account) any misconceptions that you had about the science. You don't have to go into great detail here, but do give all the facts that entered into your decision process during the analysis. List the specific questions that your I posed as well as the questions that you answered. Highlight discrepancies between the two categories of questions.

3. **Methods**:

   (a) **Source of the Data**: Describe the source and sampling methods for the data, if known. Describe the variables that are available and their meaning for the analysis. Highlight patterns of missing data as well as possible confounding by measured or unmeasured variables. This should not be a detailed presentation of descriptive statistics, however. That will come under Results.

   (b) **Statistical Methods**: Describe the methods used for the analysis at two levels. 1) Give a low-level technical description of the analysis for a potential reader. Include references for non-standard techniques. You may want to describe the software used, and certainly want to describe the methods used for assessing the appropriateness of your models. 2) Explain the basic philosophy behind the analysis techniques and how you came up with your model. For this exam, actually writing out your final model is good. Provide interpretations for all parameter estimates. Motivate transformations. Describe the use of P values and confidence intervals if they play an important role in your analysis. Explain why you didn't use more common techniques if necessary.

4. **Results**: Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done unless they provide insight into the question. Do lead the reader up to the analyses gradually.

   (a) Start off with descriptive statistics. This is an area often given short shrift in previous years. The goal is to describe the basic characteristics of the sample used to address the question, as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that needed to be addressed in the modeling, try to present descriptive statistics illustrating those issues. The basic idea is to presage all the issues you will talk about when presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.

   (b) Then go to the major models used to answer the primary questions. Present summaries of the statistical inference obtained from these models (point estimates, CI, P values). Highlight any particular issues that materially affected the models used to answer the question (confounding, interactions, nonlinearities, etc.) Tables can often be used to good effect here.

   (c) Leave secondary analyses for last and highlight the exploratory nature of those analyses. Present the results of your analyses in tables and publishing quality figures, pointing out any differences in the conclusion reached in the exploratory analyses vs. those reached in a priori analyses.

DO NOT INCLUDE OUTPUT FROM STATISTICAL PROGRAMS. (Such means little to me and nothing to a reader). When possible, use words instead of cryptic variable names. Use forms of estimates that have some meaning to a statistically naive researcher. Thus, present odds ratios rather than logistic regression parameters. Present confidence intervals rather than the values of Z, t, F, or χ^2 statistics.

5. **Discussion**: Discuss the conclusions which you feel can be drawn from the analyses. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses.

6. **Appendix**: Anything of an overly technical nature should be put in an appendix. You may want to include extensive tables in an appendix instead of the main results section.

The major theme of the above is to write to the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to anyone.

Also, lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can often be summarized in a single sentence ("We found no evidence to suggest that the final model did not fit the data adequately.") You are reporting your major results and impressions of the data. If the reader wanted to see every detail, he/she would have to do the analysis himself/herself.