Survival Analysis

HOMEWORK II

General Guidelines: although you may work in groups on this homework assignment, you must write up your own final assignment. Copying another student's homework or output is not allowed, and each student must run their own programs. In general, computer output and code will not be reviewed unless it is specifically requested as part of the assignment, and the appropriate portions of any output should be inserted directly into your homework solution (rather than attached output at the back). However students may find it useful to keep a copy of their code/output for their own documentation.

HOMEWORK II contains four questions.

Mortality Study in Myocardial Infarction Patients

The questions on this homework concern a subset of data from the Worcester Heart Attack Study. The goal of this study was to evaluate factors associated with long-term survival after an acute myocardial infarction (MI) in residents of Worcester, Massachusetts. The entire study includes over 11,000 subjects, but we will use a random subset of 500 participants.

The primary endpoint for this study was time to death (dthtime). The contents of the dataset are shown below in table 1.

The data for the study can be found on the course website as a raw dataset (MIstudy.dat; used for R/Python), as a SAS dataset (MIstudy.sas7bdat), and as a Stata dataset (MIstudy.dta). There is also a SAS program which can be used to read in the raw dataset if you prefer to create it that way.

Order	Variable Name	Description
1	id	ID number
2	age	Patient age (yrs)
3	gender	Sex (0=male, 1=female)
4	hr	Heart rate
5	sysbp	Systolic BP
6	diasbp	Diastolic BP
7	bmi	BMI (kg/m^2)
8	cvd	History of cardiovascular disease (0=no, 1=yes)
9	chf	Congestive heart failure (0=no, 1=yes)
10	av3	Complete heart block (0=no, 1=yes)
11	miord	MI order (0=first, 1=recurrent)
12	mitype	MI Type (0=non Q-wave, Q-wave)
13	year	Year of cohort (1=1997, 2=1999, 3=2001)
14	los	Length of hospital stay (days)
15	dstat	Hospital discharge status (0=alive, 1=dead)
16	afib	Atrial fibrillation (0=no, 1=yes)
17	shock	Cardiogenic shock (0=no, 1=yes)
18	dthtime	Time to death or censoring (months)
19	dthstat	Death status (1=yes, 0=censored)
20	obese	Obese (BMI>30)
21	overweight	Overweight (BMI $25-29.9$)
22	obese_ovwt	Obese/Overweight (BMI>25)

Table 1: Variables in MIstudy data set

1. Logrank and Score Tests for MI Study

The questions of interest below relate to the effect of body mass index (BMI) on mortality in the study above. We have already seen in the article by Kenchaiah et al in NEJM (2002) that obesity is a risk factor for heart failure, but recent research suggests the relationship between BMI and mortality may be more complicated. For example, in a recent article appearing in the American Journal of Epidemiology, Zheng et al (2013) found that overweight adults who maintained a stable weight after age 50 had the lower mortality risk than other weight trajectory categories, and the BMI trajectory showed a stronger association with mortality than initial BMI status.

(a) Produce a plot showing the estimated Kaplan-Meier survival functions for the endpoint of death for those who are obese or overweight (BMI$\geq$25), as compared to those of normal weight (which we will define here as BMI$<$25). What percent of the 500 patients are either overweight or obese? What do you notice about the censoring pattern?

(b) Calculate both a logrank test and Wilcoxon test of the effect of being overweight/obese on the risk of death, using either the CMH approach or a linear rank test (state which approach you have used). Which of these two tests yields a larger test statistic? Could you have predicted that based on the KM plot from (a)?

(c) Calculate the Fleming-Harrington test statistic for comparing survival distributions under several different combinations of p and q:

i. Setting p and q both to 0

ii. Setting p to 1 and q to 0

iii. Setting p and q both to 1

How do these compare to the logrank tests and Wilcoxon tests in (b)? When would you expect them to be more or less powerful than the logrank test in (b)?

(d) Ignoring the times of death, conduct a test of whether the proportions of deaths during follow-up differ for those who are overweight/obese versus those of normal weight (using a logistic regres-

sion model or based on analysis of a 2x2 table). Does this test yield similar conclusions to the tests in parts (b) and (c)? When would you expect the survival analysis to be more powerful than the comparison of proportions of deaths in the two subgroups?

2. Cox Model for Myocardial Infarction Study

Now, we will fit some Cox models to the same dataset described in **Question 1**.

(a) Fit a Cox proportional hazards model to evaluate the association of being overweight/obese with survival time, with 'obese_ovwt' as the only covariate (i.e. unadjusted). Use the discrete option for ties. Provide the Wald, Score, and LR tests for the comparison of survival distributions after MI for overweight/obese vs normal weight. Are any of these test statistics the same as either the logrank or Wilcoxon test statistics from 1(b)?

(b) Now fit an adjusted Cox proportional hazards model for evaluating the effect of 'obese_ovwt', adjusting for the effects of age, gender, systolic blood pressure (sysbp), and type of MI (mitype). Summarize the effect of overweight/obesity on survival of MI patients using both the unadjusted and adjusted hazard ratios and 95% confidence intervals. Write a short (1-2 sentence) interpretation of the HR for overweight/obesity on mortality making sure to indicate both the magnitude and direction of effect (protective or adverse effect), and whether adjustment for other covariates makes any difference.

(c) Compare the following test statistics for the effect of overweight/obesity on the risk of death by fitting the following models (don't worry about adjusting for any other covariates):

i. Test for 'obese_ovwt' from Cox PH model, stratifying by gender.

ii. Test for 'obese_ovwt' from Cox PH model, controlling for gender.

iii. Logrank test for 'obese_ovwt', stratifying by gender.

For each part, state which test statistic you use (Score, LR, Wald). How do the test statistics from (i)-(iii) compare? Do we gain any power by controlling for gender rather than stratifying by it?

(d) Briefly comment on any advantages or disadvantages of stratifying by gender rather than controlling for gender in part (c) above. Are there any additional analyses you would suggest doing to check whether the stratified analysis is appropriate?

3. Model Interpretation - Myocardial Infarction Study

Table 2 below shows the results of fitting one particular multivariable Cox proportional hazards model for time to death using some of the covariates described above. You should be able to answer the questions below using only the provided output (eg., without fitting any additional models).

Variable Name	**Estimate**	**s.e.**	**P-value**
Age	0.0500	0.0066	< 0.0001
Heart rate	0.0112	0.0029	0.0001
Diastolic BP	-0.0107	0.0035	0.0024
Sex (0=male, 1=female)	-0.2732	0.1442	0.0581
Congestive heart failure	0.7816	0.1469	< 0.0001
BMI	-0.0453	0.0163	0.0055

Table 2: Coefficient Estimate Table of Multivariable Model

(a) Write out the estimated hazard for death ($\lambda(t, Z)$) in the form of a Cox proportional hazards model, using the output from fitting the model to supply the parameter estimates.

(b) How is the "baseline" group (i.e., with hazard $\lambda_0(t)$) defined in this model in terms of all of the covariates? Does the baseline group correspond to any of the actual observations in the dataset?

(c) What is the estimated hazard ratio (HR) for death associated with a BMI of 30 as compared to a BMI of 24, holding all other covariates constant? Also provide a 95% confidence interval for the HR for a BMI of 30 vs BMI of 24, adjusting for the other covariates, and interpret.

(d) What is the estimated hazard ratio of death for a subject aged 60 years versus a subject aged 50 years, holding all other variables constant? Construct a 95% confidence interval for this estimated HR (comparing age 60 vs 50).

(e) Based on this model, calculate the estimated hazard ratio of death for a female subject aged 60 with BMI=24 versus a male subject aged 50 with BMI=30.

(f) Based on this model, describe the association of sex (male or female) with risk of death.

4. Impact of Ties on Cox Model Estimation and Testing

   Investigate the effect of tied survival times in the MI study by calculating an overall test statistic for the effect of gender on survival using at least 3 options for ties in SAS or other software. You can either fit a model for gender alone, or adjusted for other covariates; just be sure to be consistent across the different options for handling ties.

   Does it seem that ties are an important issue in this study? Please support your answer through test statistics, parameter estimates, or both as appropriate.