University of Vermont
Department of Mathematics and Statistics
Stat 229
Logistic Regression/Survival Analysis
Spring 2017

Survival Analysis Final Problem

Due in Class April 25th.

The data set to be used, also emailed to you, is:

Stat 229 Framingham Survival Problem.out.

The time variable and associated censoring indicator variable are: deathyrs and death.

Background

The Framingham Heart Study Longitudinal Data Documentation

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1948 and 5,209 subjects were initially enrolled in the study. Participants have been examined biennially since the inception of the study and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. Clinic examination data has included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, Echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of Angina Pectoris, Myocardial Infarction, Heart Failure, and Cerebrovascular disease.

The data set we are using for the exercises is a randomly selected data set from a data set provided to me by Susanne May, which is subset of the data collected as part of the Framingham study. Not all variables have been kept for use in the problem.

I purposely over sampled the events so we would be able to do meaningful analyses with only n = 500.

Table 1: Code sheet for the Stat 229 Framingham Study data used in the survival analysis problem.

Variable	Description	Units
ID	Unique identification number for each participant	1 - 500
SEX	Participant sex	0 = male, 1 = female
AGE	Age at exam	years
SYSBP	Systolic Blood Pressure (mean of last two of three measurements)	mmHg
CURSMOKE	Current cigarette smoking at exam	0=Not current smoker 1=Current smoker
CIGPDAY	Number of cigarettes smoked each day	
BMI	Body Mass Index,	weight in kilograms/height meters squared
ANYCHD	Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease	0 = no, 1 = yes
CVD	Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease	0 = no, 1 = yes
DEATH	Death from any cause	0 = no, 1 = yes
DEATHYRS	Number of years from Baseline exam to death if occurring during followup or Number of years from Baseline to censor date. Censor date may be end of followup, or last known contact date if subject is lost to followup	years

Table 2 Preliminary Main Effects Survival Time Model for the Stat 229 Framingham Study data.

```
Cox regression -- no ties

No. of subjects =          500                  Number of obs    =        500
No. of failures =          171
Time at risk    =  10294.72415
                                                LR chi2(8)       =      88.54
Log likelihood  =   -985.34127                  Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.1850352   .1632729    -1.13   0.257    -.5050441    .1349738
         age |   .0017177   .0103257     0.17   0.868    -.0185203    .0219556
       sysbp |   .0112643    .003557     3.17   0.002     .0042926     .018236
    cursmoke |   .1213079   .2649321     0.46   0.647    -.3979494    .6405652
     cigpday |   .0033698   .0102246     0.33   0.742      -.01667    .0234096
         bmi |  -.0310964   .0197571    -1.57   0.116    -.0698197    .0076269
         cvd |    .942567   .2284388     4.13   0.000     .4948352    1.390299
      anychd |   .3619552   .2160373     1.68   0.094    -.0614702    .7853805
------------------------------------------------------------------------------
```

1. Are age and sysbp linear in the log-hazard? Examine this hypothesis using fractional polynomials. If significant transformation(s) is/are found examine them graphically for clinical plausibility using the method used for logistic regression where, now, death is the binary variable. See slide 167 of the logistic regression course notes. Then replace age and/or sysbp with transformed versions and refit the model in Table 1. Even though a number of modeling steps have not been performed treat this model as your final model.
2. Assess the model's adherence to the proportional hazards assumptions.
3. Use the influence measures to examine for influential subjects.
4. Assess the model's goodness of fit using the May-Hosmer "decile of risk" test.
5. Current smoker and cigarettes per day provide and example of a problem we covered in logistic regression on slides 169 and 170. Provide estimates of the odds ratio of a non-smoker to a 20/day smoker and a 30 per day smoker to 20 per day smoker. Provide 95% confidence intervals for both odds ratios.
6. Provide estimated hazard ratios, with 95% confidence intervals, for all other model covariates.
7. Graph the estimated survival functions for cvd at the median risk score for the combined other covariates in the model.