BIOS 6641: Causal Analytics in Public Health

Data Analysis Project

Date due: Tuesday, May 16, 2017

In this project, you will be analyzing data from the Surveillance, Epidemiology and End Results (SEER) Program, a premier source for cancer statistics in the United States. SEER collects information on incidence, survival, and prevalence from specific geographic areas representing 26 percent of the US population and compiles reports on all of these plus cancer mortality for the entire US. This site is intended for anyone interested in US cancer statistics or cancer surveillance methods.

The data for the project can be downloaded from Canvas. This document and the data are located in a Final Project folder on the class webpage. A list of the variables and their meaning is given at the end of this handout.

Our time of interest is on time to relapse of lymphoma-related death. A very nice description of leukemia can be found at Wikipedia (http://en.wikipedia.org/wiki/Lymphoma). Briefly, this is a type of cancer that begins in the white blood cells and show up in the lymph nodes. There are two major categories of lymphoma, Hodgkin's lymphoma and non-Hodgkin's lymphoma. Once diagnosed with lymphoma, patients generally have treatment by radiation and/or surgery.

The goal of this project is to determine if there is a benefit on survival to surgery in conjunction with radiation relative to radiation only. You can use any methods from the course, or you can attempt to apply alternative methods. It will be important to look at descriptive statistics as much as possible.

Please contact me if you have any questions about what any of the variables mean. If you wish to analyze an alternative dataset, please set up an appointment with me to discuss the details of your dataset.

Reg: SEER Registry/Location (only 3 registries considered here: 1,2, and 20)

BYr: Year of Birth

Age: Age in years at diagnosis

Sex: gender (M = male, F = female)

Grad: grade of lymphoma (1-9; higher values indicate
more pathologically aggressive lymphoma)

Rad: radiation therapy after cancer diagnosis (0 = no, 1 = yes)

RdSrg: radiation sequence with surgery (0 = no, 1 = yes)

Dth: dead/alive (1 = dead, 0 = alive)

Site: Hodgkin's or non-Hodgkin (1 = Hodgkin, 0 = non-Hodgkin Lymphoma)

RacB: race (1 = white, 2 = black)

Stag: SEER stage (1=Localized, 2=Regional, 4=Distant, 9=Unknown)

Tim: Follow-up time in days

Caus: cause of death (0 = alive, 1 = death from lymphoma,
2 = death from non-lymphoma cancer, 3 = death from non-cancer cause)