Name:\_Rahul Avadhoot\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Last 4 Digits of Student ID Number:\_1787\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Introduction to R for Data Analysis in the Health Sciences**

**BIOST 509**

**Homework Assignment #2**

**Due: 9:00 AM on October 11, 2017**

Instructions

Enter your numerical and/or written answers to the questions into this Word file or create your own Word document (.doc or .docx) or pdf file with your answers. Your work will be graded as 1 (complete) if at least 70% of the questions are answered correctly (and/or an obvious “good faith” effort), or 0 (incomplete) otherwise. For this homework assignment, also include the R commands that you used to obtain your answers.

All questions for this homework relate to a prostate specific antigen (PSA) data set for a prospective sample of prostate cancer patients who received some sort of hormonal treatment of their prostatic carcinoma. The “psa.txt” file can be found on the Canvas web page by clicking on the “Files” link and then accessing the “Datasets” folder.

The dataset is also available on the web at <http://faculty.washington.edu/tathornt/Biost509/DataSets/psa.txt>

Additional documentation of this study of is available on the Canvas web page in the “PSA.doc” file in the “Datasets” folder.

**DESCRIPTION OF THE DATA**

The file psa.txt is a text file containing data on 50 men having hormonally treated prostate cancer. The first line of the file contains the following variable names separated by white space (spaces and/or tabs). Each successive line contains data pertinent to one of the 50 study subjects. Missing data are denoted with “NA”. Data are available for the following measurements, with individual measurements on a line separated by white space.

*ptid* = patient identifier

*nadirpsa* = lowest PSA value attained post therapy (ng/ml)

*pretxpsa* = PSA value prior to therapy (ng/ml)

*ps* = performance status (0= worst, 100= best)

*bss* = bone scan score (1= least disease, 3= most)

*grade* = tumor grade (1= least aggressive, 3= most)

*age* = patient's age (years)

obstime=time observed in remission (months)

inrem= indicator whether patient still in remission at last follow-up (“yes” or “no”)

**QUESTIONS**

1. How many individuals have missing *pretxpsa* values, i.e., PSA values prior to therapy? What proportion of individuals have missing *pretxpsa* values? Provide the patient identifier number, *ptid*, for all individuals with missing *pretxpsa* values.

Individuals have missing *pretxpsa* values = 7

Proportion of individuals have missing *pretxpsa* values = 14%

ptid for missing *pretxpsa* values = 7,14,17,34,42,45,50

R Commands:

length(which(is.na(psa$pretxpsa)))

length(which(is.na(psa$pretxpsa)))/length(psa$pretxpsa)

psa[is.na(psa$pretxpsa),]

1. What is the mean of the *pretxpsa* values for individuals with non-missing data for this variable.

Mean = 670.7512 ng/ml

R Command:

mean(psa$pretxpsa,na.rm = TRUE)

1. What proportion of individuals have censored data, i.e., the time to relapse was not observed? ***[Note: Individuals who were still in remission at last follow-up have censored data].*** What is the minimum observed time in remission among the censored observations?

Proportion of individuals having censored data = 28%

Minimum observed time = 24 months

R Commands:

nrow(psa[psa$inrem == 'yes',])/nrow(psa)

min(psa[psa$inrem == 'yes',]$obstime)

1. What proportion of individuals relapsed within the first 24 months of hormone treatment?

Proportion of individuals relapsed within 24 months = 44%

R Command:

nrow(psa[psa$inrem == 'no' & psa$obstime < 24,])/nrow(psa)

1. What is the mean nadir PSA level for individuals who were in remission at 24 months? What is the mean nadir PSA level for individuals who were out of remission before 24 months?

Mean nadir PSA levels who were in remission at 24 months = 0.5285714

Mean nadir PSA levels who were out of remission before 24 months = 31.94091

R Commands:

mean(psa[psa$inrem == 'yes' & psa$obstime > 24,]$nadirpsa)

mean(psa[psa$inrem == 'no' & psa$obstime < 24,]$nadirpsa)

1. What is the mean tumor grade for individuals in remission at 24 months? What is the mean tumor grade for individuals who were out of remission prior to 24 months?

Mean tumor grade for individuals in remission at 24 months = 2.230769

Mean tumor grade for individuals out of remission prior to 24 months = 2.235294

R Commands:

mean(psa[psa$inrem == 'yes' & psa$obstime >= 24,]$grade, na.rm = TRUE)

mean(psa[psa$inrem == 'no' & psa$obstime < 24,]$grade, na.rm = TRUE)

1. What is the mean age of individuals who were in remission at 24 months? What is the mean age of individuals who were out of remission prior to 24 months?

Mean age of individuals who were in remission at 24 months = 68.14286

Mean age of individuals who were out of remission prior to 24 months = 68.36364

R Commands:

mean(psa[psa$inrem == 'yes' & psa$obstime >= 24,]$age)

mean(psa[psa$inrem == 'no' & psa$obstime < 24,]$age)

1. What is the mean bone scan score for individuals in remission at 24 months? What is the mean bone scan score for individuals who were out of remission prior to 24 months?

Mean bone scan score for individuals in remission at 24 months = 2.071429

Mean bone scan score for individuals out of remission prior to 24 months = 2.8

R Commands:

mean(psa[psa$inrem == 'yes' & psa$obstime >= 24,]$bss)

mean(psa[psa$inrem == 'no' & psa$obstime < 24,]$bss, na.rm = TRUE)

1. For the subset of individuals who were out of remission prior to 24 months, provide a 3x3 table containing the counts (number of individuals) for each bone scan and tumor grade combination. Provide a similar table for the subset of individuals who were in remission at 24 months.

Individuals out of remission prior to 24 months:

grade

bss 1 2 3

2 0 1 2

3 3 6 4

Individuals in remission at 24 months:

grade

bss 1 2 3

1 0 2 2

2 1 2 1

3 1 2 2

R Commands:

with(psa[psa$inrem == 'no' & psa$obstime <24,],table(bss,grade))

with(psa[psa$inrem == 'yes' & psa$obstime >= 24,],table(bss,grade))