Biost 517: Applied Biostatistics I
Biost 514: Biostatistics I
Winter 2019

Homework #8
Due: Monday December 9 2019 by 9:00 AM

Written problems: To be submitted as a pdf or MS-Word compatible file via the canvas course website.

*On this (as all homeworks) R code and unedited R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the R output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

Unless explicitly told otherwise in the statement of the problem, in all problems requesting "statistical analyses" (either descriptive or inferential), you should present both

- ***Methods:** A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE R CODE.*
- ***Inference**: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details on Canvas in the "Supplementary Material" Folder.*
-

This homework uses the same dataset on a sample of generally healthy elderly subjects from four U.S. communities from the previous three homework assignments. In this homework, we are interested in assessing associations with the risk and odds of death, time to death, and creatinine levels. The data can be found on the Canvas web page by clicking on the "Files" link and then accessing the "Datasets" folder. The file "mri.txt" contains the data and the documentation is in the file "mri.pdf".

Questions:

1. Perform a logistic regression analysis to evaluate an association between baseline serum creatinine level and 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have "high" or "low" creatinine levels, where serum creatinine levels greater than 1.2 are considered to be "high", (i.e., "high" corresponds to creatinine > 1.2 mg/dl and "low" corresponds to creatinine ≤ 1.2 mg/dl). In your logistic regression model, use **an indicator of death within 5 years as the response**, and an **indicator of high serum creatinine level as the predictor**. (Only provide a formal report of inference when asked to.)
    a. Is this a saturated regression model? Explain your answer.
    b. Provide an interpretation of the slope and the intercept in your regression model in terms of the response variable (indicator of death within 5 years) and the predictor variable (high creatinine).

c. From the logistic regression model, what is the estimated odds of dying within 5 years for subjects with low creatinine levels. What is the estimated probability of dying within 5 years from the logistic regression model for subjects with low creatinine levels?

d. From the logistic regression model, what is the estimated odds of dying within 5 years for subjects with high creatinine levels. What is the estimated probability of dying within 5 years from the logistic regression model for subjects with high creatinine levels?

e. Give full inference for an association between 5-year all-cause mortality and serum creatinine levels from the logistic regression model with an indicator of death within 5 years as the response and indicator of high creatinine level as the predictor.

f. How would your answers to part b change if you were instead asked to fit a logistic regression model with indicator of death within 5 years as the response variable, but with **indicator of low serum creatinine level as the predictor**? Would the statistical evidence for an association between 5-year all-cause mortality and serum creatinine levels change? Briefly explain.

g. How would your answers to part b change if you were instead asked to fit a logistic regression model with **indicator of surviving at least 5 years as the response variable** and indicator of high creatinine level as the predictor? Would the statistical evidence for an association between 5-year all-cause mortality and serum creatinine levels change? Briefly explain.

2. In question 1, a prospective association analysis was conducted where we investigated differences in the distribution of death within 5 years across groups defined by serum creatinine level at baseline. In this question, you will now conduct a retrospective analysis and fit a logistic regression model for the distribution of serum creatinine across groups defined by vital status at 5 years. In your retrospective logistic regression model, use **an indicator of high serum creatinine level as the response**, and **indicator of death within 5 years as the predictor**. (Only provide a formal report of inference when asked to.)

    a. Provide an interpretation of the slope and the intercept in your regression model in terms of the response variable (indicator of high creatinine level) and the predictor variable (indicator of death within 5 years).

    b. From the logistic regression model, what is the estimated odds of high creatinine level for subjects who die within 5 years? What is the estimated probability of having high serum creatinine for subjects who die within 5 years?

    c. From the logistic regression model, what is the estimated odds of having a high creatinine level for subjects who survive at least 5 years? What is the estimated probability of having a high serum creatinine for subjects who survive at least 5 years?

    d. Give full inference regarding an association between 5-year all-cause mortality and serum creatinine levels from the logistic regression model with indicator of high serum creatinine as the response and an indicator of death within 5 years as the predictor.

    e. Compare the association results in part 2d from the retrospective logistic model to the association results in part 1e from the prospective logistic regression model. Briefly describe any similarities or differences.

3. Perform a regression analysis evaluating an association between serum creatinine level and 5-year all-cause mortality using the **risk difference** (RD: difference in risk or probability) of death within 5 years between groups with "high" and "low" creatinine levels (as defined in question 1). In your regression model, use **an indicator of death within 5 years as the response**, and an **indicator of high serum creatinine level as the predictor**. (Only provide a formal report of inference when asked to.)
    a. Provide an interpretation of the intercept in the regression model.
    b. Provide an interpretation of the slope in your regression model.
    c. Give full inference for an association between 5-year all-cause mortality and serum creatinine levels based on risk differences from your regression model.
4. Perform the regression analyses below comparing the distribution of death within 5 years across groups defined by the continuous measure of serum creatinine levels (i.e., do not use a dichotomized variable for serum creatinine levels for this analysis). (Only provide a formal report of inference when asked to.)
    a. Perform a regression analysis evaluating an association between 5-year all-cause mortality and serum creatinine levels using odds ratio (OR: ratio of odds) as a contrast measure.
        i. Give an interpretation of the intercept and slope of your regression model, and comment on the usefulness of each of the regression parameters scientifically.
        ii. Give full statistical inference for an association between 5-year all-cause mortality and serum creatinine levels based on odds ratios.
    b. Perform a regression analysis evaluating an association between 5-year all-cause mortality and serum creatinine levels using risk difference as a contrast measure.
        i. Give an interpretation of the intercept and slope of your regression model, and comment on the usefulness of each of regression parameters scientifically.
        ii. Is the estimated risk of 5-year all-cause mortality from your regression analysis well defined for the entire range of the observed serum creatinine levels in the data?
        iii. Give full statistical inference for an association between 5-year all-cause mortality and serum creatinine levels based on risk differences.
    c. Using the regression parameter estimates from each of the two regression models in (a) and (b), provide estimates of the risk (or probability) of death within five years for subjects that have serum creatinine levels of 0.8, 1.8, 2.8, and 3.8.

5. Time to death is a right censored quantitative variable. The analyses conducted in problems 1 – 4 used a dichotomized time to death variable according to death within five years based on the earliest censoring time of 1827 days (or slightly more than 5 years). For this problem, we will perform a survival analysis evaluating an association between time to death and serum creatinine level that appropriately accounts for right censoring of time to death.

    a. Provide a figure with Kaplan-Meier estimated survival functions for the "high" and "low" creatinine level groups, where "high" and "low" creatinine levels are defined in question 1. The two Kaplan-Meier curves for the "high" and "low" creatine groups should appear on the same plot and have an appropriate legend.

    b. Briefly comment on any differences/similarity of the survival curves from (a).

    c. Provide statistical inference for an association between time to death and serum creatinine level based on the Kaplan-Meier survival estimates for the "high" and "low" creatinine groups.