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

Homework #7
Due: Wednesday November 27, 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 homework assignments. All questions in this question relate to associations between death from any cause, creatinine, and age in a sample of generally healthy elderly subjects from four U.S. communities. 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 regression analysis to evaluate an association between serum creatinine level and 5-year all-cause mortality by comparing **geometric mean creatinine levels** across groups defined by vital status at 5 years. In your analysis, allow for the possibility of heteroscedasticity across groups defined by vital status at 5 years. (Provide formal inference where asked to.)
    a. Provide a brief description of the model you fit to address the question of an association between creatinine and 5-year all-cause mortality.
    b. What is the interpretation of the intercept? Does it have a relevant scientific interpretation?
    c. What is the interpretation of the slope? Does it have a relevant scientific interpretation?

d. Using your regression model, what is an estimate of the geometric mean creatinine level for a population of subjects who die within 5 years? How does the estimate obtained from the regression model compare to the geometric mean creatinine level in the sample for subjects who die within 5 years?

e. Using your regression model, what is an estimate of the geometric mean creatinine level for a population of subjects who survive at least 5 years? How does the estimate obtained from your regression model compare to the geometric mean creatinine level in the sample for subjects who survive at least 5 years?

f. Provide full statistical inference for an association between serum creatinine and 5-year all-cause mortality using your regression model.

2. Perform a regression analysis evaluating an association between serum creatinine levels and age by comparing **geometric mean serum creatinine levels** across age groups. In your analysis, allow for the possibility of heteroscedasticity across age groups. (Provide formal inference where asked to.)

   a. Provide a scatterplot illustrating the relationship between log serum creatinine level and age, and be sure to include in the plot a regression line from a regression model with log serum creatinine as the response variable and age as the predictor variable.

   b. Provide a description of the model you fit to address the question of an association between serum creatinine and age.

   c. What is the interpretation of the intercept? Does it have a relevant scientific interpretation?

   d. What is the interpretation of the slope? Does it have a relevant scientific interpretation?

   e. Based on your regression model, what is the estimated geometric mean serum creatinine level among a population of 75-year-old subjects, 85-year-old subjects, and 95-year-old subjects.

   f. Compare your estimates of geometric mean serum creatinine level in question (e) to estimates of the (arithmetic) mean serum creatinine levels for 75, 85, and 95 year old subjects from a linear regression model with serum creatinine levels as the response and age as the predictor, e.g., the regression model from question 3 in homework 6. Briefly discuss any similarities or differences.

   g. Provide full statistical inference about an association between serum creatinine and age based on your regression model.

   h. Provide an estimate and 95% confidence interval (CI) for the percent change in geometric mean serum creatinine between two groups that differ by 10 years in age.