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

Homework #6
Due: Wednesday, November 20, 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.)*
>
> ***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 four homework assignments. In this homework, we are interested in assessing associations among the following subset of variables of interest: self-reported race (*race*), diabetes (*diabetes*), serum creatinine level (*crt*), age (*age*) and 5 year all-cause mortality (obtained by using both the *obstime* and *death* variables). 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. Suppose we are interested in assessing the evidence for an association between diabetes diagnosis (*diabetes)* and race (*race)* in our sample of healthy elderly adults. For this question, we will only consider study participants who have a indicated race group of "white" (race=1), "black" (race=2), or "Asian" (race=3), i.e., exclude all individuals who have a race of "other" (i.e., race=4) from the analysis.

a. Provide a 3 x 2 contingency table with the counts of each combination of the *race* variable (“black”, “white”, and “Asian”) and the dichotomous variable (diabetic/non-diabetic)

b. Provide a null hypothesis and the alternative hypothesis for assessing an association between diabetes and race

c. Provide a 3 x 2 contingency table with the expected counts for table 1a above under the null hypothesis of no association between diabetes and race.

d. Provide the name of an appropriate test for assessing if the observed contingency table of counts (1a) is significantly different from the expected contingency table of counts under the null hypothesis (1c)? What is the distribution of the test statistic under the null hypothesis?

e. Give **full statistical inference** for an association between diabetes and race in the sample of elderly adults.

2. Perform statistical analyses evaluating an association between serum creatinine (*crt*) and 5-year all-cause mortality by comparing mean creatinine levels across groups defined by vital status at 5 years using **linear regression**. For this problem you do not need to provide full statistical inference. Instead, just answer the following questions.

   a. Fit two separate regression analyses. In both cases, use creatinine as the response variable. In model A, use as your predictor an indicator variable that the subject died within 5 years. In model B, use as your predictor an indicator that the subject survived at least 5 years. For each of these models, tell whether the model you fit is saturated? Explain your answer.

   b. How do models A and B relate to each other?

   c. Using the regression parameter estimate from one of your models (tell which one you use), what is the estimate of the mean creatinine among a population of subjects who survive at least 5 years?

   d. Using the regression parameter estimate from one of your models (tell which one you use), what is a confidence interval for the true mean creatinine level among a population of subjects who survive at least 5 years?

   e. Using the regression parameter estimates from one of your models (tell which one you use), what is the estimate of the true mean creatinine level among a population of subjects who die within 5 years?

   f. Using the regression parameter estimates from one of your models (tell which one you use), what is a confidence interval for the true mean creatinine level among a population of subjects who die within 5 years?

   g. Provide an interpretation of the intercept and slope from the regression model A.

   h. Using the regression parameter estimates, what are the point estimate, the estimated standard error of the point estimate, the 95% confidence interval for the true difference in means between a population that survives at least 5 years and a population that dies within 5 years? What is the P value testing the hypothesis that the two populations have the same mean creatinine level? What conclusions do you reach about a statistically significant association between serum creatinine level and 5-year all-cause mortality?

3. Perform a **linear regression** analysis evaluating an association between serum creatinine level (*crt*) and age (*age*) by comparing the distribution of creatinine level across groups defined by age as a continuous variable. (Provide formal inference where asked to.)

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

   b. Provide a scatterplot illustrating the relationship between serum creatinine level and age and include in the plot the regression line from your regression analysis.

   c. Based on your regression model, what is the estimated mean serum creatinine level among a population of 72-year-old subjects?

   d. Based on your regression model, what is the estimated mean serum creatinine level among a population of 82-year-old subjects? How does the difference between your answer to this problem and your answer to part c relate to the slope?

   e. Based on your regression model, what is the estimated mean serum creatinine level among a population of 99-year-old subjects? Do you think this estimate is a reliable estimate for mean creatinine in a population of 99-year-old subjects? Briefly explain why or why not?

   f. What is the interpretation of the intercept in your model? Does it have a relevant scientific interpretation?

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

   h. Provide full statistical inference for an association between serum creatinine level and age.

   i. Provide a 95% CI for the difference in mean creatinine across groups that differ by 5 years in age.