**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Winter 2018

**Homework #3**

Due: Friday January 26, 2018

**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 builds on the statistical analyses performed in homework #2. As such, all questions 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 statistical analysis evaluating an association between serum creatinine level and 5 year all-cause mortality by comparing **geometric mean serum creatinine levels** between groups defined by vital status at 5 years. In your analysis, allow for heteroscedasticity. Provide full statistical inference about an association between creatinine and 5 year all-cause mortality.
2. Perform a regression analysis evaluating an association between serum creatinine level and age by comparing **mean serum creatinine levels** across groups defined by age as a continuous variable. In your analysis, allow for heteroscedasticity. (Provide formal inference where asked to.)
   1. Provide a brief description of the statistical methods for the model you fit to address the question of an association between creatinine and age.
   2. Is this a saturated model? Explain your answer.
   3. Provide a scatterplot illustrating the relationship between serum creatinine and age and include in the plot the regression line from your regression analysis.
   4. Based on your regression model, what is the estimated mean creatinine level among a population of 70 year old subjects?
   5. Based on your regression model, what is the estimated mean creatinine level among a population of 85 year old subjects? How does the difference between your answer to this problem and your answer to part d relate to the slope?
   6. Based on your regression model, what is the estimated mean creatinine level among a population of 101 year old subjects? Do you think this estimate is a reliable estimate for the mean creatinine of a population of 101 year old subjects? Briefly explain why or why not?
   7. What is the interpretation of the intercept in your model? Does it have a relevant scientific interpretation?
   8. What is the interpretation of the slope?
   9. Provide full statistical inference about an association between serum creatinine and age based on your regression model.
   10. Suppose we wanted an estimate and a 95% CI for the difference in mean creatinine across groups that differ by 10 years in age. What would you report?
3. Now perform a regression analysis evaluating an association between serum creatinine levels and age by comparing the **geometric mean of serum creatinine levels** across groups defined by age as a continuous variable. In your analysis, allow for heteroscedasticity. (Provide formal inference where asked to.)
   1. Provide a description of the statistical methods for the model you fit to address the question of an association between serum creatinine and age.
   2. Based on your regression model, what is the estimated geometric mean serum creatinine level among a population of 70 year old subjects, 80 year old subjects, and 90 year old subjects.
   3. What is the interpretation of the intercept? Does it have a relevant scientific interpretation?
   4. What is the interpretation of the slope?
   5. Provide full statistical inference about an association between serum creatinine and age based on your regression model.
   6. 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.
   7. Compare your estimates of geometric mean serum creatinine level in question (b) to estimates of (arithmetic) mean serum creatinine levels for 70, and 80, and 90 year old subject from a linear regression model with serum creatinine levels as the response and age as the predictor, e.g., the regression model for problem 2 above. Briefly discuss any similarities or differences.