**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #1**

January 5, 2015

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, January 12, 2015. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata 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 Stata OR 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.***

*Keys to past homeworks from quarters that I taught Biost 517 (e.g. HW #8 from 2012) or Biost 518 (e.g., HW #1 from 2014 or HWs #1, 3 from 2008) or Biost 536 (e.g. HW #3 from 2013) might be consulted for the presentation of inferential results. Note that the requirement to provide a paragraph describing your statistical methods was new last year, and thus keys prior to 2014 do not give explicit examples of a separate paragraph. However, many past keys provide this information as an introductory sentence.*

All questions relate to associations between death from any cause and serum C reactive protein (CRP) levels in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine inflammatory biomarkers and mortality. The data can be found on the class web page (follow the link to Datasets) in the file labeled inflamm.txt. Documentation is in the file inflamm.pdf. The data is in free-field format, and can be read into R by

read.table("http://www.emersonstatistics.com/datasets/inflamm.txt",header=T)

It can be read into Stata using the following code in a .do file.

infile id site age male bkrace smoker estrogen prevdis diab2 bmi ///

aai cholest crp fib ttodth death cvddth ///

using http://www.emersonstatistics.com/datasets/inflamm.txt

Note that the first line of the text file contains the variable names, and will thus be converted to missing values. Similarly, there is some missing data recorded as ‘NA’, and those, too, will be converted to missing values. If you do not want to see all the warning messages, you can use the “quietly” prefix. You may want to go ahead and drop the first case using “drop in 1”, because it is just missing values.

Recommendations for risk of cardiovascular disease according to serum CRP levels are as follows (taken from the Mayo Clinic website):

|  |  |
| --- | --- |
| Below 1 mg/L | Low risk of heart disease |
| 1 - 3 mg/L | Average risk of heart disease |
| Above 3 mg/L | High risk of heart disease |

1. The observations of time to death in this data are subject to (right) censoring. Nevertheless, problems 2 – 6 ask you to dichotomize the time to death according to death within 4 years of study enrolment or death after 4 years. Why is this valid? Provide descriptive statistics that support your answer.

**The minimum time of follow-up among censored observations is 1,480 days, which is just over 4 years. So the vital status of everyone is known at 4 years. Therefore such dichotomization is valid.**

1. Provide a suitable descriptive statistical analysis for selected variables in this dataset as might be presented in Table 1 of a manuscript exploring the association between serum CRP and 4 year all-cause mortality in the medical literature. In addition to the two variables of primary interest, you may restrict attention to age, sex, BMI, smoking history, cholesterol, and prior history of cardiovascular disease.

**Methods: I created an indicator variable for death within 4 years of study enrollment. The variable “prevdis” was used as an indicator variable for the prior history of cardiovascular disease. The descriptive statistics are presented within groups defined by serum CRP levels (below 1 mg/L, 1 and 3 mg/L, and above 3 mg/L). I also present the descriptive statistics for the entire sample for comparison. In each group, I include the mean, standard deviation, minimum and maximum (in the format: mean (sd; min-max)) for the continuous variables (age, bmi, cholest), and I present percentages for indicator variables (sex, indicator of smoker and indicator of prior history of cardiovascular disease).**

**Results: The subjects with CRP values missing are omitted, so 4933 instead of 5000 observations are involved in my analysis. Among the group “<1 mg/L”, 1 subject was missing data on “smoker”, 1 subject was missing on “cholest”; among the “1-3 mg/L” group, 12 subjects were missing data on “BMI”, 5 subjects were missing on “smoker”; among the “>3 mg/L” group, 1 subject was missing data on “BMI”, 1 subject was missing data on “cholest”. These subjects were omitted when we calculate the descriptive statistics in each subcase. We cannot assess the impact of omitting the missing values on the generalizability of our result.**

**Of the 4933 subjects with available measurements, 428 had CRP levels lower than 1 mg/L, 3330 had CRP levels between 1 to 3 mg/L inclusive, and 1175 subjects had CRP levels greater than 3 mg/L. The following table gives the descriptive statistics within these tables. Subjects with CRP levels in lowest interval are more likely to be male than in other intervals. Subjects with the highest levels of serum CRP appeared to have higher BMI values, higher smoker rate, higher prior cardiovascular disease rate, and higher mortality rate. No consistent trend was observed across groups in age and cholesterol.**

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing mean CRP values across groups defined by vital status at 4 years.

**Methods: I compared the mean serum CRP levels of the 484 subjects who died within 4 years and the mean serum CRP levels of the 4449 subjects who survived at least 4 years. I used t-test allowing unequal variances to test the difference in the means and provided a 95% confidence interval for the difference in population means.**

**Results: The mean serum CRP for people who survived at least 4 years is 3.42 mg/L and the mean serum CRP for people who died within 4 years is 5.38 mg/L. The mean difference is about -1.96 mg/L and the 95% confidence interval with allowance for unequal variances is [-2.70, -1.21]. It means that 1.96 mg/L lower mean serum CRP among subjects who survived at least 4 years would not be unusual if the true difference in population means falls between 2.7 mg/L and 1.21 mg/L lower mean serum CRP among people who survived at least 4 years. The two-sided p-value is less than 0.001, so we can reject the null hypothesis that the mean CRP values are not different by vital status at 4 years at a 0.05 level of significance in favor of a hypothesis that death within 4 year is associated with the higher mean CRP level.**

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing geometric mean CRP values across groups defined by vital status at 4 years. (Note that there are some measurements of CRP that are reported as zeroes. Make clear how you handle these measurements.)

**Methods: Before making log transformation to the CRP values, I replaced 0 by 0.5 for those observations with CRP=0. Geometric mean of CRP levels were compared between subjects who died within 4 years and subjects who survived a least 4 years. Difference between the mean of log transformed CRP levels was tested using t-test allowing unequal variances. 95% confidence intervals for the difference in population means for log transformed CRP level were also obtained by t-test without assuming the same variance. Then I exponentiated the estimates and confidence intervals so as to get the inference on the geometric mean.**

**Results: The geometric mean serum CRP for 4449 people who survived at least 4 years is 2.03 mg/L and the geometric mean serum CRP for 484 people who died within 4 years is 2.97 mg/L. Based on the 95% confidence interval computed with an allowance for unequal variances, the observed tendency of 46.4% higher geometric mean among subjects dying within 4 years would not be unusual if the true ration of the population geometric means falls between 33.2% and 60.9% higher geometric mean among subjects dying within 4 years. The p-value is less than 0.0001 for the t test on the log transformed serum CRP levels allowing unequal variances, so we can with high confidence reject the null hypothesis that the geometric mean CRP values are not different by vital status at 4 years at a 0.05 level of significance in favor of a hypothesis that death within 4 year is associated with the higher geometric mean CRP level.**

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the probability of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods: The proportions of the subjects dying within 4 years were compared between subjects who had serum CRP greater than 3 mg/L and subjects who had serum CRP less than or equal to 3 mg/L. I used Pearson’s chi-squared test for independence to test the difference between the probabilities of death within 4 years. Then I used Wald statistics to compute the 95% confidence interval for the difference in population probability of death within 4 years.**

**Results: For the 1175 subjects with CRP > 3 mg/L, the probability of death within 4 years is 15.6%. For the 3758 subject with CRP less than or equal to 3 mg/L, the probability of death within 4 years is 8%. There is 7.6% absolute higher survival probability in subjects with CRP >3 mg/L. The 95% confidence interval for the difference in probabilities of death is [5.3%, 9.9%]. It means that with 95% confidence, the true population difference in probabilities of death falls 5.3% absolute higher survival probability to 9.9% absolute higher survival probability in subjects with CRP > 3 mg/L compared to the subjects with CRP less than or equal to 3 mg/L. By using a chi-squared test, this observation is statistically significant at a significance level of 0.05 because the two-sided p-value is less than 0.0001. Therefore we can reject the null hypothesis that the 4 year mortality rates are not associated with the serum CRP levels.**

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the odds of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods: The odds of the subjects dying within 4 years were compared between subjects who had serum CRP greater than 3 mg/L and subjects who had serum CRP less than or equal to 3 mg/L. I used Fisher’s exact test to test if the odds ratio is different from 1. The 95% confidence interval for the odds ratio was computed by the exact method.**

**Results: For the 1175 subjects with CRP > 3 mg/L, the odds of death within 4 years is 0.184. For the 3758 subject with CRP less than or equal to 3 mg/L, the odds of death within 4 years is 0.087. The observed odds ratio is 2.12. The 95% confidence interval for the odds ratio is [1.73, 2.59]. It means that with 95% confidence, the true odds ratio falls between 1.73 and 2.59. By using Fisher’s exact test, this observation is statistically significant at a significance level of 0.05 because the two-sided p-value is less than 0.0001. Therefore we can reject the null hypothesis that the odds of 4 year mortality are not associated with the serum CRP levels.**

1. Perform a statistical analysis evaluating an association between serum CRP and all-cause mortality over the entire period of observation of these subjects by comparing the instantaneous risk of death across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods: I used Kaplan-Meier estimates to estimate the survival probabilities for the two groups defined by whether the subjects have CRP > 3mg/L or not. I also presented the K-M curve on purpose for visualization. I used Cox proportional hazards regression to get the hazard ratio. I used the function in R which is “coxph(…, robust=T)” to get the Huber-White estimate for standard error, and based on that I computed the 95% confidence interval.**

**Results: As the plot and the table show, the survival probability of the 3758 subjects with CRP less than or equal to 3 mg/L is greater than the survival probability of those 1175 subjects with CRP greater than 3 mg/L in every point. The observed hazard ratio is 1.687 and the 95% confidence interval is [1.486, 1.915]. It means that with 95% confidence, the true hazard ratio falls between 1.486 and 1.915. The p-value is less than 0.0001, indicating that we can with high confidence reject the null hypothesis that the survival probability is not associated with serum CRP levels.**

1. Supposing I had not been so redundant (in a scientifically inappropriate manner) and so prescriptive about methods of detecting an association, what analysis would you have preferred *a priori* in order to answer the question about an association between mortality and serum CRP? Why?

**There are several points we should take into consideration about analyzing the data:**

1. **For some of the problems above, we dichotomized a measurement that is continuous; it would be statistically more precise to avoid doing so.**
2. **When we did log transformation to the CRP levels, we noticed that some of the values were 0. The way I did in that problem was to assign 0.5 to replace those 0’s. But I could not guarantee that 0.5 is the best value to assign and I could not assess the impact of doing so on the analysis.**
3. **The uses of geometric mean, odds ratio and hazard ratio may be hard to explain or to be understood in some scenarios. I may could try other simpler test, such as a comparison of arithmetic means, though it may not have as much precision as the tests of geometric means, etc.**
4. **It is more scientifically pleasing to condition on CRP levels and to summarize the survival distributions.**
5. **The tests and methods we apply to analysis should be what we well understand.**

**Therefore I may consider to use a test that compares means across groups instead of using odds ratio or geometric means. I would not do log transformation to this data set, and I would keep the time variable and CRP levels continuous instead of dichotomizing them. I may do proportional hazard regression on CRP levels.**