**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #1**

January 5, 2015

**TOTAL SCORE: 68**

**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 ///

systBP 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. 5/5 points

Ans: The minimal observation time of censored data is 1480 days, which is just above 4 years.

Then for this dataset, whether one subject dies before or after 4 years can be known explicitly.

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. Table layout 4/4 points; Descriptive stats 2/3 points; Summary of results 3/3 points = 9/10 points

Ans:

Methods: Indicator variable was created for death within 4 years of study enrollment. Descriptive statistics are presented within groups defined by death within 4 years or after 4 years, and also for the entire sample population. We include the mean, standard deviation, min and max for continuous variables (age, bmi, serum cholesterol measure and serum CRP levels) and include frequencies for binary variables (indicators of male, smoker, previous prevalent atherosclerotic disease).

Results: We have 5000 subjects in total, of which 6 subjects have missing value on indicator of smoker, 13 subjects have missing value on body mass index, 47 subjects have missing value on serum cholesterol level and 67 subjects have missing value on serum CRP levels. Those subjects with missing values on the variables of interest for this analysis were omitted from all analysis. We should note that these missing values might have impact on the generalizability of our results.

Of the 4911 subjects without missing data on any variables of interest for this analysis, 482 died within 4 years of study enrollment and 4429 were still alive 4 years after study enrollment. The following table presents descriptive statistics within these groups. Subjects dying within 4 years were more likely to be male, more likely to be smoker, tended to be older, tended to have higher prevalence of atherosclerotic disease before the study enrollment and tended to have lower serum cholesterol level than subjects surviving for at least 4 years after the study enrollment. Also subjects dying within 4 years tended to have higher blood CRP level: mean blood CRP was 5.39 mg/l in those observed to die within 4 years compared to a mean blood CRP of 3.42 mg/l in those surviving at least 4 years.

|  |  |
| --- | --- |
|  | Vital Status at 4 Years Post Study Enrollment |
|  | Alive at 4 Years(n=4429) | Death w/in 4 Years(n=482) | All Subjects(n=4911) |
| Age (yrs)1 | 72.4 (5.30; 65-98) | 76.3 (6.71; 65-100) | 72.8 (5.57; 65-100) |
| Male (%) | 40.08 | 60.17 | 42.05 |
| Smoker (%) | 11.94 | 14.32 | 12.18 |
| Prior prevalent atherosclerotic disease (%) | 20.89 | 41.91 | 22.95 |
| BMI (kg/m2) 1 | 26.7 (4.70; 14.70-58.80) | 26.3 (5.00; 14.80-48.10) | 26.7 (4.72; 14.7-58.8) |
| Serum Cholesterol (mg/dl) 1 | 212.5 (38.91; 78.00-430.0) | 204.1 (41.47; 73.00-396.0) | 211.7 (39.24; 73.00-430.0) |
| Blood CRP (mg/l) 1 | 3.42 (5.86; 0.00-108) | 5.39 (8.11; 0.00-55.0) | 3.61 (6.15; 0.00-108) |

1 Descriptive statistics presented are the mean (standard deviation; min-max)

Table did not include a footnote or column specifying missing values. (- 1 point)

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.

Ans: 10/10 points

Methods: Mean CRP levels were compared between subjects who died within 4 years of study enrollments and those who survived more than 4 years. Difference in mean CRP was tested using a two- sample t test allowing for unequal variance. 95% confidence intervals for the difference in mean CRP were similarly based on that same handling of variance.

Results: Mean CRP was 5.39 mg/l among 482 subjects dying within 4 years of study enrollment and was 3.42 mg/l among 4429 subjects surviving more than 4 years after study enrollment. Mean CRP of subjects dying within 4 years tends to be 1.97 mg/l higher than those who survived more than 4 years after the study enrollment. The observed data is not unusual if the true difference of population mean CRP was between 1.22 mg/l and 2.72 mg/l higher among subjects dying within 4 years of study enrollment compared to those surviving at least 4 years after study enrollment. Also this observation is statistically significant at a 0.05 level of significance (two-sided P<0.001), we can reject the null hypothesis that the mean CRP levels are not different by vital status at 4 years in favor of a hypothesis that death within 4 years is associated with higher mean CRP.

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.)

Ans: 9/10 points 🡪 Minus 1 point for not specifying how CRP values of zero were handled.

Methods: Geometric mean of CRP levels were compared between subjects who died within 4 years of study enrollments and those who survived more than 4 years after study enrollment. Difference in the mean of log transformed serum CRP was tested using a t test allowing for unequal variance. 95% confidence intervals for the difference in population means of log transformed CRP were similarly based on that same handling of variance. Estimates and confidence interval were then exponentiated in order to obtain inference on the geometric mean.

Results: Geometric mean of CRP was 2.98 mg/l among 482 subjects dying within 4 years of study enrollment and was 2.03 mg/l among 4429 subjects surviving more than 4 years after study enrollment. Geometric mean of CRP of subjects dying within 4 years tends to be 46.8% higher than those who survived more than 4 years after the study enrollment. The observed data is not unusual if the true difference of population mean CRP was between 33.5% and 61.4% higher among subjects dying within 4 years of study enrollment compared to those surviving at least 4 years after study enrollment. Also this observation is statistically significant at a 0.05 level of significance (two-sided P<0.001), we can reject the null hypothesis that the geometric means of CRP levels are not different by vital status at 4 years in favor of a hypothesis that death within 4 years is associated with higher geometric mean of CRP.

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).

Ans: 9/10 points 🡪 Minus 1 point for slight discrepancies in the numerical results compared to the answer key.

Methods: Indicator variable was generated as indicator of high CRP and used for the following analysis. The probabilities of death within 4 years of study enrollment were compared between subjects with high CRP (> 3 mg/l) and those with low to average CRP (≤ 3 mg/l). Differences in the probability of death within 4 years of study enrollment were tested using Pearson's chi squared test. 95% confidence intervals for the difference in population mortality of death within 4 years were computed using Wald statistics.

Results: Among the 1172 subjects with high CRP (> 3 mg/l), 15.61% were observed to die within 4 years of the study enrollment, while 8.00% of 3739 subjects with low to average CRP (≤ 3 mg/l) were observed to die within 4 years of the study enrollment. Subjects with high CRP tended to have 7.62% absolute higher probability of death within 4 years of the study enrollment. The observed data is not unusual if the true difference in survival probabilities were between a 5.36% and a 9.87% higher absolute probability of survival in the high CRP group compared to the low to average CRP group. Using a chi-squared test, this observation is statistically significant at a 0.05 level of significance (two-sided P<0.001), and we can reject the null hypothesis that the probabilities of death within 4 years of the study enrollment are not associated with serum CRP.

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).

Ans: 10/10 points

Methods: The odds of subjects dying within 4 years of study enrollment were compared between subjects with high CRP and subjects with low to average CRP level. An odds ratio different from 1 was tested using Chi-squared test. 95% confidence intervals for the odds ratio were computed using Cornfield's methods.

Results: Of the 1172 subjects with high CRP (>3 mg/l), the odds of dying within 4 years from study enrollment was 0.185, while for the subjects with low to average CRP (≤ 3 mg/l), the odds of 4 year mortality was 0.0869. Based on a 95% confidence interval, this observed odds ratio of 2.13 for the comparison of the high CRP group to the low to average CRP group was not unusual if the true odds ratio were between 1.75 and 2.59. A chi squared test two-sided p value (P<0.001) suggests that we can reject the null hypothesis that the odds of 4-year mortality are not associated with 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).

Ans: 8/10 points 🡪 Minus 1 point for copying the exact wording from HW 1 key and minus 1 point for incorrect values in the table.

Methods: The survival distribution was estimated using Kaplan-Meier estimates with strata defined by serum CRP greater than 3 mg/l and less than or equal to 3 mg/L. Difference in survival distributions between those two groups was tested using the logrank statistic. The hazard ratio and 95% CI was computed using Cox proportional hazards regression with the Huber-White sandwich estimator of the standard errors.

Results: The following graph and table depicts Kaplan-Meier estimates of survival probability for the subjects with high CRP and those with low to average CRP level. Apparent from that graph is the tendency for higher instantaneous risk of death in subjects with high CRP at every point during the study. The instantaneous risk of death is estimated to be 69% higher for the high CRP group compared to the low to average CRP group. Based on a 95% confidence interval, this observed hazard ratio of 1.69 for the comparison of the high CRP group to the low to average CRP group would not be judged unusual if the true hazard ratio were between 1.48 and 1.92. A logrank test two-sided p value less than 0.001 suggests that we can with high confidence reject the null hypothesis that probability of survival is not associated with serum LDL levels.



|  |  |
| --- | --- |
|  | Kaplan-Meier survival probabilities |
|  | High CRP (> 3 mg/l) | Low to average CRP (≤ 3 mg/l) |
| 2 years | 0.926 | 0.971 |
| 4 years | 0.844 | 0.920 |
| 6 years | 0.753 | 0.852 |
| 8 years | 0.647 | 0.768 |
| Values in the table are incorrect (see key). |

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? 8/10 points

Ans: I prefer comparing the mean of serum CRP level across groups defined by vital status, which can be achieved by using two-sample t test allowing for unequal variance. First, vital status is a natural categorical variable. In contrast, serum CRP level is a continuous variable. Dichotomization of serum CRP level may lead to artificial results. Second, mean is a sensitive detector of many changes in the distribution of variables. Also mean is easily understood and widely accepted. Lastly, we have very convenient and easy ways to compare means (like t test).