**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 ///

systBP aai cholest crp fib ttodth death cvddth ///

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

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.

**Notice that the minimum follow up time for censored subjects is 1480 days, which is more than 4\*365= 1460 days. Thus, no one in the study has been censored within 4 years of study enrollment so that we know whether a subject died or not within 4 years study enrollment. Therefore censoring is not an issue here and it is valid to dichotomize.**

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: All subjects are divided into two groups based on whether she or he survived within 4 years (1460 days). For two groups and overall subjects, descriptive statistics of serum CRP, age, sex, BMI, smoking history, cholesterol, and prior history of cardiovascular disease (atherosclerotic disease). For continuous variables, including serum CRP, age, BMI and cholesterol, mean, standard deviation, minimum and maximum are presented; for discreet variables, including sex, smoking history, atherosclerotic disease, percentages are present. If any of observation has missing value in one of these variables above, such observation(s) will be omitted from the analysis, assuming comparing to the total number subjects number of omitted patients are relatively very small.**

**Results: Total number of subjects in the study is 5000. Among those 5000 subjects, 67 subjects have missing value in blood C reactive protein; 6 subjects have missing value in smoking history; 13 subjects have missing value in BMI; 47 subjects have missing value in cholesterol. After omitting these subjects, 4911 subjects were included in this descriptive analysis as well as following inferential analyses (89 subjects omitted, 13 deaths within 4 years since study enrollment).**

**Among 4911 subjects, 4429 were still alive at years since study enrollment and 482 died within 4 years. Based on the following descriptive statistics, subjects who died within 4 years after study enrollment tended to be elder male with lower serum cholesterol measurement, smoking history and atherosclerotic disease history. The mean of blood c reactive protein of subjects died within 4 years was 5.39 mg/L, which was higher than the mean of blood c reactive protein of subjects were still alive after 4 years, 3.42 mg/L.**

|  |  |
| --- | --- |
|  | Vital Status at 4 Years Since Study Enrollment |
|  | Alive at 4 Years (n=4429) | Death within 4 Years (n=482) | All subjects (n=4911) |
| Male | 40% | 60% | 42% |
| Age (yr)\* | 72.42 (5.30, 65 - 98) | 76.3 (6.71, 65 - 100) | 72.8 (5.57, 65 - 100) |
| BMI (kg/m^2)\* | 26.7 (4.69, 14.7 - 58.8) | 26.3 (4.98, 14.8 - 48.1) | 26.7 (4.72, 14.7 - 58.8) |
| Cholesterol (mg/dl)\* | 212.5 (38.9, 78 - 430) | 204.1 (41.5, 73 - 396) | 211.7 (39.2, 73 - 430) |
| Blood C reactive protein (mg/L)\* | 3.42 (5.86, 0 - 108) | 5.39 (8.11, 0 - 55) | 3.61 (6.15, 0 - 108) |
| Smoking History | 12% | 14% | 12% |
| Atherosclerotic Disease | 21% | 42% | 23% |

\*descriptive statistics presented are the mean (standard deviation, minimum - maximum)

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 use t-test, allowing possibility of unequal variance to test differences in mean CRP values across groups defined by vital status at 4 year. Noticed that Satterthwaite method is used to approximate variances in test and confidence interval construction.

Results: The mean CRP was 5.39 mg/L among 482 subjects who died within 4 years since study enrollment; the mean CRP was 3.42 mg/L among 4429 subjects were still alive after 4 years. The mean CRP among subjects died within 4 years was 1.97 mg/L higher than the other group and, with 95% confidence interval and allowing unequal variance; this mean difference would not be usual if the true population mean difference was between 1.23 mg/L and 2.72 mg/L higher in among subjects who die within 4 years. At a 0.05 level significance (two-side p-value < 0.0001, allowing unequal variance), we can reject the null hypothesis that the mean CRP values are not different between subjects who die within 4 years and subjects who are still alive at years. And the results in favor of the hypothesis that serum CRP is associated with vital status at 4 years.

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: I use t-test, allowing possibility of unequal variance to test differences in mean of log transformed CRP values across groups defined by vital status at 4 year. Later, the estimates and confidence interval were exponentiated. Noticed that Satterthwaite method is used to approximate variances in test and confidence interval construction. Also, for 0 values in CRP, 0.5 was added to avoid the undefined log operation.

Results: The geometric mean CRP was 2.98 mg/L among 482 subjects who died within 4 years since study enrollment; the geometric mean CRP was 2.03 mg/L among 4429 subjects were still alive after 4 years. The geometric mean CRP among subjects died within 4 years was 47% higher than the other group and, with 95% confidence interval and allowing unequal variance this geometric mean difference would not be usual if the true population geometric mean ratio was between 1.33 and 1.61 in among subjects who die within 4 years. At a 0.05 level significance (two-side p-value < 0.0001, allowing unequal variance), we can reject the null hypothesis that the geometric mean CRP values are not different between subjects who die within 4 years and subjects who are still alive at years. And the results in favor of the hypothesis that serum CRP is associated with vital status at 4 years.

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: I used Pearson's chi squared test to test the independence of 4 year death probabilities between subjects have serum CRP higher than 3 mg/L and subject with equal or lower than 3 mg/L serum CRP. 95% confidence interval for the death probabilities difference was constructed using Wald statistics.

Results: 1172 subjects have high serum CRP and 15.6% of them died within 4 years. 3739 subjects have low serum CRP and 8.0% of them died within 4 years. Based on 95% confidence interval, 7.6% higher death probability for subject who has higher serum CRP would not be unusual if the true death within 4 years probability differences in the population is between 5.3% and 9.9% higher in high serum CRP subjects. At a 0.05 level significance (two-side p-value < 0.0001), we can reject the null hypothesis that the death probability within 4 year is independent of high serum CRP status (> 3 mg/L) and the result is in favor of the hypothesis that the death probability within 4 years is associated with the high serum CRP status.

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: I used Pearson's chi squared test to test the independence of odds of 4 year death between subjects have serum CRP higher than 3 mg/L and subject with equal or lower than 3 mg/L serum CRP. 95% confidence interval for odds of death difference was constructed using Wald statistics.

Results: 1172 subjects have high serum CRP and the odds of dying within 4 years was 0.185. 3739 subjects have low serum the odds of dying within 4 years was 0.087. Based on 95% confidence interval, the odds ratio between vital status of two groups is 2.129 which would not be unusual if the true odds ratio in the population is between 1.748 and 2.593. At a 0.05 level significance (two-side p-value < 0.0001), we can reject the null hypothesis that the odds of dying within 4 year is independent of high serum CRP status (> 3 mg/L) and the result is in favor of the hypothesis that the odds of dying within 4 years is associated with the high serum CRP status.

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: Two Kaplan-Meier survival curves were constructed to estimate survival distributions between subjects with high CRP (> 3 mg/L) and rest of total subject. Logrank test was used to test the difference in two survival distributions. Cox proportional hazards regression estimated hazard ratio and 95% CI.

Results: The Kaplan-Meier curves shown that subjects with high serum CRP (n=1172) have lower survival probability than subjects with lower serum CRP (n=3739). At a 0.05 level significance, the logrank test rejected the null hypothesis survival distributions between two groups are the same (two-side p-value<0.0001) and thus the results in favor for the hypothesis that the survival probability is associated with high CRP status. The estimated instantaneous risk of death is 69% higher for subjects with high CRP than subjects with lower CRP. With 95% confidence, the hazard ratio of 1.69 would not be unusual if the true odds ratio is between 1.49 and 1.92. Therefore, the data is in favor of the hypothesis that the subjects with high CRP status are associated with higher instantaneous risk of death.

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?

I will prefer the analysis from question 3, comparing mean CRP values across groups defined by vital status at 4 years by t-test. Firstly, I chose arithmetic mean over geometric mean because it is much earlier to interpret for most of people. Although it might be scientifically meaningfully to consider geometric mean, without a prior knowledge, I do not have any reason to choose geometric mean. Secondly, I prefer not to dichotomize CRP. Although there is a clinical reference, dichotomization on a continuous variable by 3 mg/L is still arbitrary and it will lose information. Also, odds is not easy to interpret. Therefore, I prefer not to choose analysis from question 6 to 9.