**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

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.

Answer: Following is the descriptive statistics of observation time for the censored patients.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | N | Mean | Std | Min | Max |
| Censored time | 3879 | 2603 | 414 | 1480 | 2942 |

As shown in the table above, the first censored subject occurs on 1480 days (equal to 4.05 years). Therefore, dichotomizing the data according to death within or after 4 years 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: Patients are divided into three groups based on their serum CRP levels (less than or equal to 1mg/L, between 1and 3 mg/L, and greater than or equal to 3 mg/L). Descriptive statistics are presented in the table below of each group as well as the entire sample. For continuous variables (age, BMI and cholesterol level), the mean, standard deviation and the data range are included. For binary variables (sex, smoking status, prior history of CVD and 4 year all-cause mortality), the data is presented in percentages.

Results: This set of data includes a total of 5000 subjects, but there are 67 subjects are missing data on their CRP levels and they are omitted in the data analysis. The rest of the subjects (n=4933) are divided into three groups based on their serum CRP levels (less than or equal to 1mg/L, between 1and 3 mg/L, and greater than or equal to 3 mg/L). There are 13 subjects who are missing data on their BMI; 3 subjects who are missing data on their cholesterol level; 6 subjects who are missing data on their smoking status. But they are not of the main interest of this study, so these subjects are still included.

The descriptive statistics within these groups are demonstrated in the table below. There are no obvious differences across groups in sex, age, BMI and cholesterol level. However, in the group with high level of CRP (≥ 3mg/l), there are higher percentage of subjects with smoking status (15.7%) and prior cardiovascular disease (27.2%) compared to other two groups. The group of subjects with serum CRP level greater than or equal to 3mg/l appeared to have a higher mortality rate (13.9%) compared to other two groups (6.6% for the group with CRP ≤ 1mg/L and 8.6% for the group with CRP level between 1and 3 mg/L).

|  |  |
| --- | --- |
|  | Blood C reactive protein |
|  | ≤1(mg/l)(n=1969) | 1-3(mg/l)(n=1088) | ≥3(mg/l)(n=1876) | Total(n=4933) |
| Male(%) | 44.8% | 44.4% | 37.7% | 42.0% |
| Age(years) | 73.1(5.7, 65-98) | 72.5(5.4, 65-100) | 72.6(5.5, 65-93) | 72.8(5.6, 65-100) |
| BMI  | 25.1(3.8, 14.7-43.2) | 26.9(4.5, 15.1-53.2) | 28.1(5.2, 15.3-58.8) | 26.7(4.7, 14.7-58.8) |
| Cholesterol (mg/dl) | 210.3(38.0, 73-407) | 213.9(40.4, 78-354) | 211.8(39.7, 96- 430) | 211.7(39.2, 73-430) |
| Smoking (%) | 9.3% | 11.3% | 15.7% | 12.2% |
| Prior CVD (%) | 18.9% | 22.7% | 27.2% | 22.9% |
| Death with 4 years | 6.6% | 8.6% | 13.9% | 9.8% |

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: A t test (with the assumption of unequal variances) was performed to compare mean serum CRP levels between subjects who died within 4 years and those who survived after 4 years. 95% confidence intervals for the difference in population means were estimated based on the same assumption.

Results: For the subjects who survived at least 4 years after study enrollment (n=4449), the mean serum CRP is 3.4 mg/L. For the subjects who died within 4 years (n=484), the mean serum CPR is 5.4 mg/L. On average, the serum CRP level of the subjects who died with 4 years is 2.0 mg/l higher than that of the subjects who died after 4 years. With 95% confidence, this observed result would not be unusual if the true mean serum CRP level of the subjects who died within 4 years is higher by 1.2 to 2.7mg/l than that of the subjects who died after 4 years. This observation is statistically significant at a 0.05 level of significance based on a t test (two-sided P< 0.0001). Therefore, we can reject the null hypothesis that the mean serum CRP levels are not different between the subjects who died within 4 years and the subjects who died after 4 years with high confidence. We conclude that the death of the subjects within 4 years of study enrollment is associated with high mean serum 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: A t test (with the assumption of unequal variances) was performed to compare geometric mean serum CRP levels between subjects who died within 4 years and those who survived after 4 years. The subjects with the serum CRP level equal to 0 mg/L have been changed to 0.5mg/L (half of the detection limit in this study). 95% confidence intervals for the difference in population geometric means were estimated based on the same assumption and then exponentiated to obtain inference.

Results: There are a total of 428 subjects with the serum CRP reading of 0mg/L. In order to log transfer the data, these values have been replaced by 0.5mg/L (half of the detection limit in this study). For the subjects who survived at least 4 years after study enrollment (n=4449), the geometric mean serum CRP is 2.03 mg/L. For the subjects who died within 4 years (n=484), the mean serum CPR is 2.97 mg/L. On average, the geometric mean serum CRP level for subjects who died within 4 years is 46.4% higher than that of subjects who survived at least 4 years of enrollment. With 95% confidence, the observed data would not be unusual if the true geometric mean CRP level for subjects who died within 4 years is 33.2% to 60.9% higher than that of subjects who survived at least 4 years. This observation is statistically significant at a 0.05 level of significance based on a t test (two-sided P< 0.0001). Therefore, with high confidence we can reject the null hypothesis that the geometric mean serum CRP levels are not different between the subjects who died within 4 years and the subjects who died after 4 years. We conclude that the death of the subjects within 4 years of study enrollment is associated with high geometric mean serum 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).

Method: The subjects dying within 4 years of study enrollment were divided into two group based on their serum CRP level (less than 3mg/L; greater than or equal to 3 mg/L). The probability of death was compared between groups using Pearson’s chi squared test for independence. 95% confidence intervals for the difference between groups in population were also estimated.

Results: For the group whose serum CRP level was less than or equal to 3 mg/L(n=3758), 8.0% of the subjects died within 4 years of study enrollment. For the group whose serum CRP level was greater than 3 mg/l (n=1175), 15.6% of the subjects died within 4 years. On average, the survival probability of the group with high CRP level is lower by 7.6% compared to the group with low CRP level. With 95% confidence, this observed result would not be unusual if the true survival probability of the group with CRP level greater than 3mg/L is lower by 5.3% to 9.8% than the group with CRP level less or equal to 3mg/l. This observation is statistically significant at a 0.05 level of significance based on a chi squared test (two-sided P< 0.0001). In summary, we can reject the null hypothesis with high confidence and conclude that the survival probabilities within 4 years of study enrollment are associated with 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 subjects dying within 4 years of study enrollment were divided into two group based on their serum CRP level (less than 3mg/L; greater than or equal to 3 mg/L). The odds of death were compared between groups using Pearson’s chi squared test. 95% confidence intervals for the odds ratio were also estimated.

Results: For the group whose serum CRP level was less than or equal to 3 mg/L, the odds of dying within 5 years from study enrollment was 0.087. For the group whose serum CRP level was greater than 3 mg/l, the odds of 5 year mortality was 0.184. The observed odds ratio is 2.12 for the comparison of the subjects of high serum CRP levels to the subjects of low serum CRP levels. With 95% confidence, this observed result would not be unusual if the true odds ratio was anywhere between 1.74 to 2.58. This observation is statistically significant at a 0.05 level of significance based on a chi squared test (two-sided P< 0.0001). In summary, we can reject the null hypothesis with high confidence and conclude that the odds of death within 4 year are 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).

Methods: The subjects were divided into two group based on their serum CRP level (less than or equal to 3mg/L; greater than 3 mg/L). Kaplan-Meier was used to estimate the survival distribution in two groups. Difference in survival distributions between two groups was tested using a logrank statistic. The hazard ratio and 95% confidence intervals were calculated using Cox proportional hazards regression.

Results: Kaplan-Meier curves in the following graph demonstrate the survival probability of the two group based on their serum CRP level (3758 subjects with serum CRP level less than or equal to 3 mg/L and 1175 subjects with serum CRP level greater than 3 mg/L). It is apparent from the graph that survival probability for group with the high CRP level is lower than the group with low CRP level. The instantaneous risk of death is estimated to be 68.7% higher for the group with high serum CRP level compared to the group with low serum CRP level. The observed hazard ratio is 1.687 for the comparison of the subjects in high serum CRP group to the subjects in the low serum CRP group. With 95% confidence, this observed result would not be unusual if the true hazard ratio was anywhere between 1.485 to 1.916. This observation is statistically significant at a 0.05 level of significance based on a logrank test (two-sided P< 0.0001). Therefore, we can reject the null hypothesis with high confidence and conclude that the survival probability is 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?

Answer: I would prefer a t test to compare the mean CRP level across groups defined by vital status at 4 years. Because t test with the assumption of unequal variances is usually valid and robust. The results of the difference of the means between groups are easy to calculate (compared to the geometric means in problem 4) and be better understood. Usually it is not recommended to dichotomize a continuous variable, but in this study, it is better to dichotomize the survival time than the serum CRP levels.

Dividing subjects into multiple groups based on their CRP levels can also be considered due to the mechanism study in biochemistry. Usually a trend can be observed between groups.