**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.

**This is a valid dichotomization because all of the subjects have either died within 4 years of enrollment or they are still alive at 4 years. If we look at the times to death, the min is 5 days and the max is 2942, or about 8 years. We also know that each person will die eventually, so we can group all of the patients who haven’t died within 4 years into the second category. Since the variable is defined as either time to death or end of study, we know that the longest any person could be in the study was 8 years, and thus dichotomizing the times to death on the scale provided is a valid step (since it is less than the max time we have all of the information we need).**

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: In order to make the table more presentable, I have broken up some of the continuous variables based on how they are reported by a clinician. Thus CRP is grouped into normal (<1 mg/L), borderline (between 1 and 3 mg/L) and high (>3 mg/L). BMI is grouped into underweight (below 18.5), normal (below 24.9), overweight (below 30) and obese (above 30). Cholesterol is grouped into desirable (less than 200), borderline (less than 240) and high (greater than 240). Age is grouped into under 85 and over 85. The others are all already binary variables.**

**RESULTS: There are 67 people with a missing serum CRP value, and thus these are not included in further analysis. Of the remaining 4933 subjects, some were missing data on some of the other variables of interest, but none were missing data on all of them. Thus we removed the missing data from analysis of the individual variable, but not from the data set. Of these 4933 subjects, 428 had <1 mg/L serum CRP, 3330 had between 1 and 3 mg/L, and 1175 had > 3 mg/L. The following table presents the results of a descriptive analysis. The patients with lowest serum CRP levels tended to be older on average, and there was a clearly increasing trend in BMI as serum CRP level increased. Those with lowest serum CRP also had the lowest cholesterol, though the other two were relatively similar. There was a decreasing trend in the proportion of male subjects as serum CRP level increased, an increasing trend in proportion who smoked, an increasing trend in proportion who had previous history of disease, and an increasing proportion who died within 4 years.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Serum CRP level** |  |  |
|  | **Normal (<1mg/L) (n=428)** | **Borderline (between 1 and 3 mg/L) (n=3330)** | **High (>3 mg/L) (n=1175)** | **Any Level (n=4933)** |
| **Age (yrs)** | 73.45 (5.80; 65-94) | 72.74 (5.524; 65-100) | 72.74 (5.581; 65-93) | 72.80 (5.565; 65-100) |
| **BMI** | 23.82 (3.639; 21.28 - 38.60) | 26.39 (4.306; 23.50 - 53.20) | 28.45 (5.463; 24.52 - 58.80) | 26.66 (4.722; 23.50 - 58.80)  |
| **Cholesterol** | 206.0 (40.53; 181.0 - 407.0) | 212.8 (38.57; 187.0 - 363.0) | 210.5 (40.39; 184.0 - 430.0) | 211.7 (39.23; 186.0 - 430.0) |
| **Male (%)** | 45.56 | 43.3 | 37.02 | 42 |
| **Smoke (%)** | 9.6 | 11.01 | 16.43 | 12.18 |
| **Previous History (%)** | 18.22 | 21.47 | 28.77 | 22.9 |
| **Died within 4 years (%)** | 4.91 | 8.41 | 15.57 | 9.811 |

**NB: Values in the first three rows of the table are given in the format Mean (SD; Min – Max)**

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: Mean CRP values were compared between those subjects who had died prior to 4 years and those who were still alive at 4 years. Differences in the mean were compared using a t-test allowing for unequal variances. 95% confidence intervals were computed.**

**RESULTS: Mean CRP level in the 4449 subjects alive at 4 years is 3.42 mg/L and the mean CRP level in the 484 subjects dead is 5.38 mg/L. This corresponds to a difference in means of -1.95, with a 95% confidence interval of [-2.70, -1.21]. Thus this data would not be unusual with a true difference in means within the interval. Based on the t-test described above, we get a p-value <0.05, and thus we can with high confidence reject the null hypothesis and conclude that the mean CRP values in the two groups are different.**

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: Now we note that there are some measurements of CRP that are reported as zeroes. Since we believe that a zero could arise from a serum CRP below the measurement threshold, we wish to include these in analysis. Thus, we add half of the minimum nonzero value in the data set, which means we add 0.5, to each zero entry for CRP in our calculations. Then we calculate the geometric means, and then we perform a t-test allowing for unequal variances with the null hypothesis that the geometric means are equal and the alternative that they are not. A 95% confidence interval is then computed.**

**RESULTS: Geometric mean serum CRP was .78 mg/L among the 4449 subjects alive at 4 years, and 1.08 mg/L among those who had died. Based on a 95% confidence interval of -.476, -.286] the observed tendencey of a 38.1% lower geometric mean serum CRP level among subjects surviving at least 4 years is not unusual. Based on the t-test described above, we get a p-value of <0.0005, and thus we reject the null hypothesis and conclude that the geometric mean CRP values in the two groups are different.**

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 proportion of subjects dying within 4 years of study enrollment were compared across those with high serum CRP (>3 mg/L) and those with low CRP (<= 3 mg/L). Differences in the probability of death were compared using a chi-squared test for independence, and a 95% confidence interval was computed.**

**RESULTS: Of the 1175 subjects whose serum CRP was above 3 mg/L, 15.6% died within 4 years. Of the remaining 3758 subjects, 8.0% died within 4 years of enrollment. Based on the chi-squared test above, we get a 95% confidence interval of [0.053, 0.099] which means that our estimated difference in proportions is consistent with a true difference in this interval. The chi-squared test returns a p-value <.0005, so we reject the null hypothesis in favor of the alternative, or that the probability of death within 4 years with high serum CRP is greater than the probability of death within 4 years with low 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).

**METHODS: We construct a 2x2 table and calculate the odds ratio. This is given by**

|  |  |  |
| --- | --- | --- |
|  | **Died** | **Alive** |
| **High** | **183** | **992** |
| **Low** | **301** | **3457** |

**The odds of dying within 4 years in the high serum CRP are then compared to the odds of dying within 4 years with low serum CRP. They are compared against a ratio of 1.**

**RESULTS: The the odds ratio is given by 2.11, and a 95% confidence interval is [1.73, 2.57]. Since this interval does not contain 1, we reject the null hypothesis and conclude that the odds of death in the high CRP group tend to be higher than the odds of death in the low CRP group.**

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 survival function in each group was estimated using Kaplan-Meier estimates. Then the differences in the survival functions were compared using the logrank statistic. We also used a cox proportional hazards regression to obtain an estimate and confidence interval for the hazard ratio.**

**RESULTS: Based on the 95% confidence interval for the hazard ratio based on the Cox Proportional Hazards regression, our estimate of a hazard ratio of 1.69 is consistent with a true hazard ratio of between 1.49 and 1.92. The logrank statistic is 66.8 with 1 degree of freedom, and a p-value of < 0.005. Thus we reject the null hypothesis and conclude that the instantaneous risk of death is different between the two groups (and is in fact higher in the high CRP group).**

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?

**If we had to decide *a priori* which method to choose, we must consider a few points. First, it we get close to answering the scientific question of interest by conditioning on serum CRP measurements and then summarizing the survival distribution, because we presume the serum CRP value to precede death. Second, it is more statistically precise to not dichotomize a continuous variable, so it is best to leave serum CRP alone. Therefore, *a priori*, I would choose to do a comparison of the means to answer the question about an association between mortality and serum CRP, since I am not incredibly comfortable with interpreting and using geometric means or hazard functions and the odds ratio has been ruled out by not dichotomizing the data.**