**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.

The minimum length of follow-up time among censored observations (death = 0) is 1480 days, which is slightly longer than 4 years. Therefore, the mortality outcome of every study participant is known at 4 years.

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:** Descriptive statistics are categorized into groups defined by CRP level (less than 1 mg/L, between 1 and 3 mg/L, and greater than 3 mg/L), as well as in the entire sample. An indicator variable was created for past history of any cardiovascular disease (CVD = angina, MI, TIA, stroke) at time of study enrollment and also for smoking history. Within each group defined by CRP level, for continuous variables (age, BMI, cholesterol) the table includes the mean, standard deviation, minimum and maximum. For binary variables (sex and indicators of prior history of CVD & smoking) percentages are given.

**Results**: Data is available for 5000 study participants. The following number of study participants had missing data for some of the variables of interest: 13 missing for BMI, 47 missing for cholesterol level, and 6 missing for smoking history. 428 participants had a CRP level less than 1 mg/L, 3330 had a CRP level between 1 and 3 mg/L, and 1242 had a CRP level greater than 3 mg/L. The following table presents descriptive statistics within these groups. Participants with a CRP level greater than 3 mg/L were slightly less likely to be male than those in other groups. There was a trend towards increasing BMI, increased history of smoking, and increased history of prior cardiovascular disease with increasing CRP level.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **C-reactive Protein (CRP) Level** | | | |
|  | < 1 mg/L  (n=428) | 1-3 mg/L  (n=3330) | >3 mg/L  (n=1242) | Total  (n=5000) |
| **Male (%)** | 45.6% | 43.3% | 37.0% | 41.9% |
| **Age (years)\*** | 73.5 (5.80; 65-94) | 72.7 (5.52; 65-100) | 72.9 (5.71; 65-93) | 72.8 (5.60; 65-100) |
| **BMI (kg/m2)\*** | 23.8 (3.64; 15.6-38.6) | 26.4 (4.31; 14.7-53.2) | 28.4 (5.48; 15.3-58.8) | 26.7 (4.74; 14.7-58.8) |
| **Cholesterol\*** | 206 (40.52; 109-407) | 213 (38.57; 73-363) | 211 (40.61; 97-430) | 212 (39.29; 73-430) |
| **History of smoking** | 9.6% | 11.0% | 15.9% | 12.1% |
| **History of CVD** | 18.2% | 21.5% | 28.7% | 23.0% |

\* Statistics presented as follows: 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:** Mean CRP levels were compared between participants who died within 4 years of study enrollment and those who survived at least 4 years. The t test that allows for the possibility of unequal variances was used to compare means between the two groups and also used to generate 95% confidence intervals for the difference in population means.

**Results**: Mean CRP level was 3.42 m/L among the 4449 participants who survived at least 4 years after study enrollment and 5.38 mg/L among the 484 participants who died within 4 years. Based on a 95% confidence interval, the finding of a 1.95 mg/L lower mean CRP level among participants dying earlier would not be unusual if the true difference in population means was anywhere between a 1.21 mg/L to 2.70 mg/L lower mean CRP level among participants who die within 4 years. Based on the results of the t test, this finding is statistically significant at a 0.05 level of significance (two-sided p<0.00001). Therefore, we can reject the null hypothesis that the mean CRP levels are not different by mortality outcome at 4 years and support the alternative hypothesis that death within 4 years is associated with higher mean CRP levels.

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.)
2. 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 between participants who had CRP levels greater than 3 mg/L and those whose CRP level was less than or equal to 3 mg/L. Differences in the probability of death within 4 years were tested using Pearson’s chi squared test for independence. 95% confidence intervals for 4 year mortality probabilities were computed using Wald statistics.

**Results**: Of the 3758 participants whose CRP level was less than or equal to 3 mg/L, 8.0% were observed to die within 4 years. Among the 1242 participants with a CRP level greater than 3 mg/L, 15.6% died within 4 years of study enrollment. Based on a 95% confidence interval, this 7.6% lower survival probability in subjects with higher CRP level would not be unusual if the true difference in survival probabilities were anywhere between a 5.4% to a 9.8% lower probability of survival in the high CRP level group compared to the low group. Using a chi squared test, this finding was statistically significant at a 0.05 level of significance (two-sided p<0.0001). Therefore, we can reject the null hypothesis that the survival probabilities are not associated with CRP levels and support the alternative hypothesis that high CRP levels are associated with lower probability of survival to 4 years.

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 participants dying within 4 years of study enrollment were compared between participants who had CRP levels greater than 3 mg/L and those whose CRP level was less than or equal to 3 mg/L. An odds ratio different from 1 was tested using Fisher’s exact test. 95% confidence intervals for the odds ratio was also computed using exact methods.

**Results**: Compared to the 3758 participants whose CRP level was less than or equal to 3 mg/L, the odds of dying within 4 years from study enrollment was higher among the 1242 participants with a CRP level greater than 3 mg/L with an odds ratio of 2.13. Based on a 95% confidence interval, this observed odds ratio of 2.13 would not be unusual if the true odds ratio were anywhere between 1.74 to 2.59. A Fisher’s exact test two-sided p value of <0.00001 suggests that we can reject the null hypothesis that the odds of 4 year mortality is not associated with CRP levels and support the alternative hypothesis that high CRP levels are associated with higher odds of mortality at 4 years.

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).
2. 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?