**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #1**

January 5, 2015

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 time of follow up of censored observations in this study is 1,480 days, which is slightly over 4 years. This means that the status of all individuals since enrollment was known up to 4 years of follow up. It is therefore valid to dichotomize the variable “time to death” into death within 4 years and those who survived beyond 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:**

An additional indicator variable “death within 4 years” was created in addition to the variables within the data set. The predictor of interest – serum blood levels of C reactive protein (CRP) was categorized into three groups; less than 1mg/L, 1-3 mg/L and above 3mg/L. Descriptive statistics that included the main outcome mortality defined as a categorical variable “death within 4 years” and population characteristics such as age, cholesterol blood levels, body mass index, sex, smoking status and prior history of cardiovascular disease were reported within each CRP subgroup and for the total sample. We reported the mean, standard deviation, minimum and maximum for continuous variables like age in years, cholesterol blood levels (mg/dl) and body mass index and percentages for binary variables such as sex (male), smoking status (Yes), prior history of cardiovascular disease (Yes) and death within 4 years (Yes).

**Results:**

The sample size for this study is 5,000, however 67 subjects had missing C reactive protein (CRP) measurements, 11 of whom were dead within 4 years. We excluded CRP missing data from all analysis, and in addition the remaining missing data for the three variables of interest such as BMI, smokers and cholesterol levels which had 13, 6 and 3 missing values respectively. Of note BMI missing data had an additional four dead within 4 years. Because of the exclusion of all missing data, we will not be able to assess its impact on the study generalizability.

Out of the 4,911 subjects analyzed, 426 (8.7%) had CRP measurements below 1mg/L, 2,615 (53.3%) between 1 and 3 mg/L and 1,870 (38.1%) above 3mg/L. Table 1 below provides descriptive statistics of demographic, medical and behavioral risk factors and death within 4 years for each CRP category. No consistent trend was seen across groups in age and cholesterol blood levels. Subjects with CRP below 1mg/L were more likely to be male. There seems to be a positive trend noted with increasing body mass index and higher CRP levels. Similarly, the prevalence of those with prior chronic vascular disease and snokers is higher with increased levels of CRP. Finally, the prevalence of death within four years is higher with increasing CRP level category suggesting a positive trend.

Table 1. Study population characteristics and mortality within the three C reactive protein serum level categories

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | C reactive protein Levels (mg/L) | | | |
|  | Below 1 mg/L  (N=426) | 1 - 3 mg/L  (N=2,615) | Above 3 mg/L  (N=1,870) | Any level  (N=4,911) |
| Demographic Profile |  |  |  |  |
| Age (years) 1 | 73.4  (5.79, 65-94) | 72.8  (5.58, 65-100) | 72.6  (5.49, 65-93) | 72.8  (5.57, 65 -100) |
| Male (%) | 45.5% | 44.5% | 37.8% | 42.0% |
| Medical Profile |  |  |  |  |
| Body Mass Index (Kg/m2) 1 | 23.8  (3.64, 15.6 -38.6) | 26.1  (4.17, 14.7-53.2) | 28.1  (5.18, 15.3- 58.8) | 26.7  (4.72, 14.7-58.8) |
| Cholesterol (mg/dl) 1 | 206.15  (40.46, 109-407) | 212.52  (38.64, 73-363) | 211.80  (39.70, 96-430) | 211.70  (39.24, 73-430) |
| Have prior history of Cardiovascular Disease (%) | 18.3% | 20.7% | 27.2% | 22.9% |
| Behavioral Profile |  |  |  |  |
| Smoker (%) | 9.6% | 10.1% | 15.7% | 12.2% |
| Mortality |  |  |  |  |
| Death within 4 years (%) | 4.9% | 7.6% | 14.0% | 9.8% |

* Statistics displ
* istic 1 Variable Statistic display include 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:** The t test that assumes unequal variances was used to test the difference in mean CRP measurements in mg/L units between subjects who died within 4 years to those who survived at least 4 years (Satterthwaite's approximation). Similarly the generated 95% confidence interval of the difference is under the assumption of unequal variances.

**Results:** The mean CRP level for those who died within 4 years (N=482) is 5.39 mg/L, while the mean of those who did not (N=4,429) is 3.42mg/L. The observed difference in mean CRP levels is 1.97 mg/L higher among those who died within 4 years compared to those who survived at least four years. This difference in mean CRP will not be unusual if the true population difference is between 1.23 and 2.72 mg/L higher among those who died within 4 years compared to those who survived at least four years. The difference in the means is also statistically significant at a 0.05 level (two sided P <0.0001), rejecting the null hypothesis that states that there is no difference between the mean levels of CRP between those who die within 4 years or after.

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:** Since some CRP values are equal to zero, 0.5 (half of lowest value) was added to all CRP values. The CRP values were then log transformed (base 10). The t test that assumes unequal variances was used to test the difference in mean of log transformed CRP measurements (mg/L) between subjects who died within 4 years and those who survived at least 4 years (Satterthwaite's approximation). Similarly the generated 95% confidence interval of the mean log difference is under the assumption of unequal variances. The geometric mean estimates were then exponentiated back to unit scale and 0.5 subtracted to make inferences in point estimates and 95% confidence intervals.

**Results:** The geometric mean CRP level for those who died within 4 years (N=482) is 3.08 mg/L, while that for those who survived beyond 4 years (N=4,429) is 2.02mg/L. The observed difference in geometric mean of CRP levels is 35.4% higher among those who died within 4 years compared to those that survived at least 4 years. This difference in log means will not be unusual if the true population difference is between 26.4% and 43.7% higher among those who died within 4 years compared to those who survived at least four years. The difference in the log means is also statistically significant at a 0.05 level (two sided P <0.0001), rejecting the null hypothesis that states that there is no difference between geometric means of CRP between those who die within 4 years or survived beyond 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:**

Two categories of CRP serum levels were created, a “high” defined as greater than 3mg/L and a “low” as equal to or less than 3. The proportion of those who died within 4 years in subjects with “high” serum CRP was compared to the proportion of subjects who died within 4 years with “low” serum CRP levels. Pearson’s chi squared test was used to test the differences in probability between both groups and 95% confidence intervals of the difference in probabilities were computed using Wald statistics.

**Results:** 14.0% of those with serum CRP greater than 3mg/L (N=1,870) died with 4 years since enrolment compared to 7.3 % of those with serum equal to or less than 3 mg/L (N=3,041). The difference in probability of death within 4 groups of 6.7% is not unusual if the true difference in probabilities between “high” and “low” CRP serum groups is between 5.3% and 9.8%. Chi square test indicates that the difference is significant at 0.05 level (two sided P<0.0001), rejecting the null hypothesis that states that there is no difference in probability of death within 4 years between those who have greater than 3mg/L versus (“high” CRP serum) to those equal to or less than 3mg/L (“low” CRP serum).

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).

**Method:** The odds of those who died within 4 years since enrollment was compared between those who had “high” serum CRP levels defined earlier as greater than 3mg/L versus “low” serum CRP levels as equal to or less than 3mg/L. Pearson’s chi squared test was used to test the differences in odds between both groups and 95% confidence intervals of the difference in odds were computed using Wald statistics.

**Results:** The odds of dying within 4 years since enrolment is 0.16 among participants (N=1,870) with “high” serum CRP defined as greater than 3mg/L. In comparison, the odds of dying within 4 years since enrolment is 0.08 in participants (N=3,758) with low serum CRP defined as equal to or less than 3 mg/L. The odds ratio between these groups is 2.07 and this finding will not be unusual if the true odds ratio is 1.71 and 2.50 (Cornfield). Chi square test indicates that the difference is significant at 0.05 level (two sided P<0.0001), rejecting the null hypothesis that states that there is no difference in odds of death within 4 years between those who have greater than 3mg/L or “high” CRP levels compared those with “low” CRP levels equal to or less than 3mg/L.

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).

**Method:** Kaplan-Meier estimates was used to determine the association for CRP serum level subgroups defined as “high” greater than 3mg/L and “low” equal to or less than 3mg/L and all cause mortality over the entire period of study. The survival distributions’ difference was tested using log rank statistic. Cox proportional hazard regression that does not assume equal variances was used to calculate the hazard ratio and 95% confidence intervals.

**Results:** Figure 1 below illustrates the survival probability for participants with “high” serum CRP levels defined as greater than 3mg/L and “low” serum CRP levels equal to or less than 3mg/L. The survival curve for those with ‘high” serum CRP levels depicted as a blue line have lower survival rates compared to those with “low” serum CRP levels depicted in red. The log rank test two sided p value is <0.0001 rejects the null hypothesis of no difference of survival curves between both groups, indicating that CRP serum levels are significantly associated with survival rates. The instantaneous risk of death is 60% higher for those with “high” CRP levels compared to “low” CRP levels. Alternatively, the hazard ratio is 1.60 and this ratio will be unusual if the true hazard ratio is between 1.42 and 1.8.

Figure 1. Kaplan-Meier Survival estimates for populations with “high” and “low” serum CRP levels over study period



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?

1. CRP is an inflammatory marker that is associated with several diseases and some types of cancer. However, the study population is elderly healthy population and CRP levels are therefore not expected to vary much or be in multiplicative scale to consider log transformation of CRP.

2. The study design is prospective with an assumption that CRP levels occur before and have an associated “effect” on mortality. I will therefore examine the mortality conditioning different levels of CRP.

3. Odds ratios and geometric measures are hard to interpret

4. I would like to utilize as much of the data as possible, i.e. include censored data

For the above reasons, my preference is to use Kaplan Meier Survival estimates that utilizes censored data but the setback is that CRP data will be categorized into groups and not utilizing all the data hence losing on precision.