Biostatistics 515/518

Homework 1

ID code: 1491

Date: 01/12/15

**Question 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:** Although time to death is a right censored variable, the following analyses will be done on data that is dichotomized on the time to death according to death within 4 years of study enrolment or death after 4 years. This is valid because the minimum time of study follow up among the censored observation is 1,480 days (just over 4 years). This means that we know whether every subject is alive or dead at 4 years of follow up time, making dichotomizing at this time frame valid.

**Method:** The minimum time of follow up was determined by performing descriptive statistics on the time to death stratified by whether or not a death was observed.

**Question 2:** 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.

**Method:** In order to provide descriptive statistics to address the question of an association between serum CRP and 4 year all-cause mortality, I chose to provide summary statistics for subjects grouped into the three levels of serum CRP. I first grouped subjects according to their Serum CRP (mg/L) values into three groups: <1 mg/L, 1-3 mg/L, and >3 mg/L. I then calculated summary descriptive statistics for the subjects in this dataset stratified by these CRP groups.

**Results:** In order to address the question of a possible association between serum CRP and 4 year all-cause mortality in the elderly, data was collected on 5000 elderly ( >65 years) Medicare patients from four different communities. Sixty seven individuals (including 12 who died during the study duration) did not have serum CRP data and have been excluded from all analyses. The subjects with serum CRP values were stratified into 3 different groups based on serum CPR levels: <1 mg/L, 1-3 mg/L, and >3 mg/L, based on the Mayo Clinic. The least number (425 428 in the key) of subjects was in the lowest CRP category, while 3,330 subjects were in the middle category, and 1,175 subjects were in the highest category. Some summary statistics of these groups can be seen in Table 1. Interestingly, across the serum CRP groups there were trends in characteristics that could impact mortality rates. Specifically BMI, percentage of subjects with a history of smoking, and percentage with prevalent cardiovascular disease at enrollment all trended higher in the higher serum CRP groups. The percentage of patients who were male in each group showed the opposite trend. In contrast, serum LDL levels and the age of the patients didn’t have any clear trends across these groups.

Table 1.

|  |  |
| --- | --- |
| Characteristics | Subject Groups by Serum CRP\* levels |
|  | < 1 mg/L(n=425 428 in key) | 1-3 mg/L(n=3330) | > 3 mg/L(n=1175) | All Subjects(n = 4933) |
| Age (yrs) | 73.5 (5.8, 65–94) | 72.7 (5.5, 65–100) | 72.7 (5.6, 65–93) | 72.8 (5.6, 65–100) |
| Male (%) | 45.6 | 43.3 | 37.0 | 42.0 |
| BMI§ | 23.8 (3.6, 15.6–38.6) | 26.4 (4.3, 14.7–53.2) | 28.5 (5.5, 15.3–58.8) | 26.7 (4.7, 14.7–58.8)  |
| Serum LDLª (mg/dL) | 206 (40.5, 109–407) | 212.8 (38.6, 73–363) | 210.5 (40.4, 97–430) | 211.7 (39.2, 73-430) |
| History of Smoking (%) | 9.6 | 11.0 | 16.4 | 12.2 |
| History of CVD‡ (%) | 18.2 | 21.5 | 28.8 | 22.9 |
| Died in <4 years (%) | 4.91 | 8.41 | 15.57 | 9.81 |

Numbers represent the mean (standard deviation, minimum – maximum), unless stated otherwise

\* CRP, C-reactive protein

§ BMI, Body mass index

ª LDL, Low density lipoprotein

‡ CVD, Cardiovascular disease

**Question 3.** 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.

**Method:** I created an indicator variable of death status at 4 years (i.e. 1461 days) to stratify the subjects with CRP data into two groups. Because the study aims to determine if subjects with elevated inflammatory markers (e.g. serum CRP levels) are at increased risk for atherosclerotic or general mortality, I used a one-sided t-test (with an alternative hypothesis that the group who had died within 4 years of observation would have greater mean serum CRP levels than the group who hadn’t) to compare the mean serum CRP levels between the two groups.

**Results:**  Subjects who died within 4 years of observation had a mean serum CRP level of 5.38 mg/L (95% CI 4.65 – 6.10 mg/L)while those who were alive at this time-point had a mean serum CRP level of 3.42 mg/L (95% CI 3.25 – 3.59 mg/L). The group who had died thus had a mean serum CRP level that was 1.95 mg/L higher than those who remained alive, which is consistent with a true difference of means between infinity and 1.33 mg/L higher for those who died (95% confidence interval). Because the one-sided p-value is <0.0001, we can reject the null hypothesis that there was no difference between the means of serum CRP levels among those stratified by vital status at 4 years.

**Question 4.** 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.)

**Method:**  Before performing a log transformation to compare geometric means, I first needed to deal with the Serum CRP levels that had been recorded as zeros. Since we are dealing with measurements in blood, it is reasonable to assume that there are non-zero levels in each subject, even if some are below the limit of detection of our tests (and thus marked zero in this dataset). I first converted all zero values to 0.5 mg/L, which was one half of the limit of detection of the test (without knowledge of which test was actually administered, I took the lowest non-zero values in the dataset (i.e. 1 mg/L) to be the lower limit of detection of the test). I then log transformed these data and did descriptive statistics as well as a one-sided t-test to compare the means between the groups created when stratifying on vital status at 4 years (see question 4 for vital status stratification and rationale for one-sided test).

**Results:** Subjects who died within 4 years of observation had a geometric mean serum CRP level of 2.97 mg/L (95% CI 2.71 – 3.25 mg/L) while those who were alive at this time-point had a geometric mean serum CRP level of 2.03 mg/L (95% CI 1.97 – 2.09 mg/L). The group who had died had a geometric mean serum CRP level that was 1.96 times higher than those who remained alive, which is consistent with a true ratio of geometric means of serum CRP levels between 1.332 and infinity times higher for those who died (95% confidence interval). Because the one-sided p-value is <0.0001, we can reject the null hypothesis that there was no difference between the geometric means of serum CRP levels among those stratified by vital status at 4 years.

**Question 5.** 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 first created an indicator variable to stratify all subjects with CRP measurements into two groups: those with serum CRP measurements greater than 3 mg/L and those with serum CRP less than or equal to 3 mg/L. In a similar argument as that given in question 3, I used Pearson’s Chi-squared test (with an alternative hypothesis that the group with high serum CRP levels would have a higher proportion of subject who had died within 4 years of observation than the group with low serum CRP levels) to compare the proportion (probability) of death within 4 years between the two groups stratified by serum CRP levels.

**Results:** The group with high serum CRP (defined as greater than 3 mg/L) had a mean probability of death within 4 years of 0.1557 (95% CI 0.1350 – 0.1765), while those with low serum CRP levels had a probability of death within 4 years of 0.0801 (95% CI 0.0714 – 0.0888). The group with high serum CRP had a 0.0757 greater probability of death than those with low serum CRP, which is consistent with a true difference in probability of death within 4 years between 1 and 0.05623 higher for the group with high serum CRP (95% confidence interval). The p value of <0.0001 indicates that we can reject the null hypothesis that the probability of death within 4 years is the same for the groups stratified by low or high serum CRP levels.

**Question 6.** 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:** In order to compare the odds of death within 4 years across groups defined by serum CRP levels, I first calculated the odds in each group. I then calculated the odds ratio between the high serum CRP group and the low serum CRP group. I used a Wald approximation with critical values from the Z distribution to calculate the 95% confidence interval for this odds ratio.

**Results:** The group with high serum CRP (defined as greater than 3 mg/L) had an odds of dying within 4 years of 0.1844, while the odds of dying within 4 years for the low serum CRP group was 0.0871. The odds ratio of death within 4 years comparing the high serum CRP group to the low serum CRP group is 2.1187, which is consistent with a true odds ratio of between 1.7402 and 2.5796 (95% confidence interval). Since this confidence interval does not cross 1.0 (i.e. no difference in odds of death within 4 years between the two groups), we can reject the null hypothesis that the odds of death within 4 years is the same between two groups stratified into low and high serum CRP levels.

**Question 7.** 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:** This association was analyzed in two main ways: graphically and through a logrank test on the survival functions for the groups stratified into low and high serum CRP levels (see above). The main step here was to use analyses that focus on the survival function to gain insight on the hazard function (i.e. the instantaneous risk of death), which is valid because there is an inherent relationship between the two. The Kaplan-Meier curves were constructed for each serum CRP group, and 95% confidence intervals were determined using complementary log log standard errors in R. To test whether the survival curves were significantly different, a logrank test was employed. Throughout these analyses, we are assuming that the censoring in this data is non-informative.

**Results:** The group with high serum CRP (defined as >3 mg/L) had consistently lower probabilities of survival across the entire observation period **(see below)**. The lower probability of survival began early during the observation time. Based on the two-sided p value of < 0.0001, we can reject the null hypothesis that there are equal survival probabilities in groups defined by serum CRP levels. Because of both the Kaplan-Meier curves and the logrank test, we can conclude that there is a higher instantaneous risk of death among the group with higher serum CRP levels.



**Question 8.** 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 have probably approached this question by doing the analysis described in question 7 (i.e. comparing instantaneous risks of death). In all other analyses, we had to restrict analyses to a specific time point, mostly limited by the nature of our data (i.e. no censoring before 4 years). This rather arbitrary decision limits the insight into *how* CRP levels impacts all-cause mortality over time. While we do have to stratify in this analysis (i.e. we create a cut-off of 3mg/L for serum CRP levels), this stratification is based on scientific recommendations (if we can believe the Mayo Clinic) and could be easily applied to each real-world patient. It is harder to know how an arbitrary “4 year survival” analysis would be applied to each and every patient.