**BIOST 518/515: Applied Biostatistics II/Biostatistics II**

Emerson, Winter 2015

**Homework 01**

January 19, 2015

*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*

*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:**

Typically, for (right) censoring data, Kaplan Meier estimates are used instead of sample descriptive statistics, because we only know the lower bounds of follow-up times for censored observations but not their exact values.

Nevertheless, problems 2-6 ask to dichotomize the time to death according to death within 4 years of study enrolment or death after 4 years. This is valid because among the 3,879 censored observations, the minimum follow-up time is 1,480 days, which is approximately 4.05 years (assuming an average of 365.25 days per year). This means that all censored observations survived at least through 4 years and dichotomizing the time to death according to death within 4 years of study enrolment and death after 4 years will not give an incorrect or biased categorization of time to death.

*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.*

**Methods:**

I provide two tables, 1a and 1b, for the selected variables in this dataset. An indicator variable was create for death within 4 years of study enrolment based on the original time-to-death variable in the dataset.

For table 1a, descriptive statistics are shown within groups defined by serum CRP level (below 1mg/L, between 1 and 3 mg/L, and above 3 mg/L) as well as for the entire sample. Descriptive statistics included for continuous variables (age, BMI, cholesterol) are (arithmetic) mean, standard deviation, minimum and maximum. Descriptive statistics included for binary variables (death within 4 years of study enrolment, sex, smoking history, prior history of cardiovascular disease) are percentages.

Table 1b gives number of observations with missing data for selected variable (rows) as well as subgroups and the entire sample (columns).

**Results:**

The total number of observations in the dataset is 5,000. None of the subjects is missing data follow-up time from enrolment to death/end of study. However, 67 of those subjects (11 who died within 4 years and 56 died after 4 years) are missing data on serum C reactive protein (CRP) level. These subjects should be omitted from all analyses. There is no additional information to determine how these omissions may impact our analysis.

There are missing data on other variables, including BMI, smoking history, and cholesterol level. Table 1b presents the number of observations with missing data on these variables by subgroups as well as all subjects (excluding the 67 observations omitted as mentioned earlier). Since we are interested in the association between serum CRP and 4-year all-cause mortality, such observations with missing data on other variables are not omitted from later analysis.

Of the 4,933 subjects with available CRP measurements, 428 had serum CRP measurements less than 1mg/L (low risk of heart disease), 3,330 had measurements between 1 and 3 mg/L (average risk), and 1,175 had measurements above 3mg/L (high risk). Table 1a shows the descriptive statistics within these subgroups and all subjects.

Not surprisingly, subjects in the low-risk group had the lowest 4-year all-cause mortality rate while subjects in the high-risk group had the highest rate. There is no apparent trend across subgroups in age. The percentage of male was slightly lower in the high-risk group compared to the other two. Subjects in the low-risk group had the lowest mean BMI and smallest BMI range, while those in the high-risk group had the highest mean BMI and largest BMI range. The proportions of smokers in the low-risk and average-risk groups are slightly different and smaller than that of the high-risk group. There is no clear trend in the cholesterol level across subgroups. The mean cholesterol levels are slightly different, with high standard deviations and wide ranges of measurements. Also unsurprisingly, the low-risk group had the smallest percentage of prior cardiovascular disease while this percentage was the highest among those in the high-risk group.

**Table 1a. Descriptive Statistics within Groups by Serum CRP Level and Sample**

|  |  |
| --- | --- |
| **Variable** | **Serum C Reactive Protein (CRP) Level** |
| **Low Risk (< 1mg/L)****(n = 428)** | **Average Risk (1 – 3mg/L)****(n = 3,330)** | **High Risk (> 3mg/L)****(n = 1,175)** | **All Subjects (any level)****(n = 4,933)** |
| **Death within 4 years (%)** | 4.9% | 8.4% | 15.6% | 10.0% |
| **Age (years)1** | 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% | 42.5% |
| **BMI1** | 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) |
| **Smoker (%)** | 9.6% | 11.0% | 16.4% | 12.2% |
| **Cholesterol (mg/dL)1** | 206.0 (40.5; 109 - 407) | 212.8 (38.6; 73 - 363) | 210.5 (40.4; 97 - 430) | 211.7 (39.3; 73 - 430) |
| **Prior cardiovascular disease (%)2** | 18.2% | 21.5% | 28.8% | 23.3% |

*1 Descriptive statistics presented are: mean (standard deviation; minimum – maximum)*

*2 Prior cardiovascular disease includes any of previous angina, MI, TIA, or stroke*

**Table 1b. Number of Observations with Missing Data by Other Variables and Subgroups**

|  |  |
| --- | --- |
|  | **Serum C Reactive Protein (CRP) Level** |
| **Variable** | **Low Risk****(< 1mg/L)****(n = 428)** | **Average Risk****(1 – 3mg/L)****(n = 3,330)** | **High Risk****(> 3mg/L)****(n = 1,175)** | **All Subjects****(any level)****(n = 4,933)** |
| **Age** | 0 | 0 | 0 | 0 |
| **Sex (male)** | 0 | 0 | 0 | 0 |
| **BMI** | 0 | 12 | 1 | 13 |
| **Smoking history (smoker)** | 1 | 5 | 0 | 6 |
| **Cholesterol** | 1 | 0 | 2 | 3 |
| **Prior cardiovascular disease** | 0 | 0 | 0 | 0 |

*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.*

**Methods:**

An indicator variable is created to dichotomized follow-up time according to death within 4 years of study enrolment or death after 4 years, assuming 365.25 days in a year. Groups defined by vital status at 4 years are subjects who died within 4 years of study enrolment and those who survived at least 4 years.

Mean serum CRP values were then compared across these two groups. Difference in the means were tested using a two-sample t-test that does not assume equal variance, i.e. allows the possibility of unequal variances across groups. A 95% confidence interval were constructed accordingly, without assuming equal variance.

**Results:**

Defined by vital status at 4 years, there were 484 patients who died within 4 years of study enrolment and 4,449 patients who survived at least 4 years since enrolment. The mean serum CRP values were 5.38 mg/L and 3.42 mg/L, respectively, across the two groups. The mean serum CRP among subjects dying within 4 years tend to be 1.95 mg/L higher than the mean serum CRP among subjects dying after 4 years. With 95% confidence, this result is consistent with the true mean serum CRP among those who died within 4 years being anywhere between 1.21 mg/L to 2.70 mg/L higher than the true mean among those who survived at least 4 years. Using a two-sample t-test that does not assume equal variances, this result is statistically significant at the 0.05 critical threshold with a two-sided p-value of less than 0.0001. As a result, we reject the null hypothesis that the mean serum CRP values are not different by vital status at 4 years, in favor of an alternative hypothesis that mortality at 4 years is associated with mean CRP.

*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.)*

**Methods:**

A log-transformation was performed on the serum CRP values. Prior to performing the log-transformation, CRP measurements of zeroes were replaced by 0.5. In general, when dealing with zeroes in log-transformation, such zeroes were replaced by half of the lowest limit of detection. In this case, we do not know the exact precision of the serum CRP measurement, but we know that the smallest value recorded after 0 was 1. In addition, only 4% (21 among 484) of patients who died within 4 years and only 9% (407 among 4,449) of patients who survived at least 4 years had serum CRP measurements of zeroes. Thus adding 0.5 to these zeroes measurements can make sure that the log-transformed values are closed enough to zeroes and not creating huge outliers in calculating the geometric mean.

Comparing geometric means is equivalent to comparing the mean of the log-transformed values. In that sense, geometric mean serum CRP values were compared between subjects who died within 4 years of study enrolment and those who survived at least 4 years since study enrolment by performing a two-sample t-test that does not assume equal variances on the means of the log-transformed CRP values across these two groups. A point estimate and a 95% confidence interval for the true difference in population means of the log CRP values across these two groups were then exponentiated back to the normal scale to obtain a point estimate and a 95% confidence interval for the geometric mean.

**Results:**

The geometric mean serum CRP value was 2.97 mg/L among the 484 patients who died within 4 years of study enrolment and 2.03 mg/L among the 4,449 patients who survived at least 4 years. The geometric mean serum CRP among subjects died within 4 years tend to be 46.39% higher than the geometric mean serum CRP among those died after 4 years. With 95% confidence, this result is consistent with the true geometric mean serum CRP among those who died within 4 years being anywhere between 33.16% and 60.93% higher the true geometric mean among those who survived at least 4 years. Using a two-sample t-test that does not assume equal variances, this result is statistically significant at the 0.05 critical threshold with a two-sided p-value of less than 0.0001. As a result, we reject the null hypothesis that the geometric mean serum CRP values are not different by vital status at 4 years, in favor of an alternative hypothesis that mortality at 4 years is associated with geometric mean serum CRP.

*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:**

Comparing the probability of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3mg/L) is equivalent to a test comparing the proportions of those who died within 4 years of study enrolment across the group of patients with high serum CRP and the group without high serum CRP. The proportions of subjects dying within 4 years of study enrolment were compared across these two groups by the Pearson’s chi squared test on the difference of the two proportions. A 95% confidence interval for the true difference in proportions was computed using Wald statistics

**Results:**

There were 1,175 patients who had serum CRP above 3mg/L (“high” serum CRP) and 3,758 patients who had serum CRP of 3mg/L or below (“normal/low” serum CRP). 15.57% of the subjects with high serum CRP died within 4 years of study enrolment, while only 8.01% of those with normal/low serum CRP died within 4 years. The proportion of dying within 4 years of study enrolment among subjects with high serum CRP tend to be higher than such proportion among subjects with normal/low serum CRP by 7.56% in absolute term. With 95% confidence, this result is consistent with such true proportion in the group with high serum CRP being anywhere between 5.32% and 9.81%, in absolute term, higher than such proportion in the group with normal/low serum CRP. Using a chi square test, this result is statistically significant at the 0.05 critical threshold with a two-sided p-value of less than 0.0001. As a result, we reject the null hypothesis that the proportions of dying within 4 years are similar across the groups defined by level of serum CRP (i.e. level of serum CRP and mortality at 4 years are independent), in favor of an alternative hypothesis that the probability of dying within 4 years is associated with serum CRP level.

*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).*

**Methods:**

The odds of death within 4 years of study enrolment among the group with high serum CRP and the group with normal/low serum CRP are compared by examining the odds of dying within 4 years of each group and the odds ratio. Table 2 presents a summary of observations by vital status at 4 years and serum CRP level. Fisher’s exact test was used to determine whether the true odds ratio was different than 1. The 95% confidence interval for the true odds ratio was constructed in a similar manner using the exact method. The expected counts were high enough to construct the 95% confidence interval using the Wald’s method, but for both convenience (the Fisher’s exact test command in R gives the confidence interval for odds ratio while there is no direct way to get such interval for Wald’s method in R) and precision, I used the Fisher’s exact method.

**Table 2. Summary of Observations by Vital Status at 4 Years and Serum CRP Level**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **High Serum CRP** | **Normal/Low Serum CRP** | **Total** |
| **Died within 4 years** | 183 | 301 | 484 |
| **Survived at least 4 years** | 992 | 3,457 | 4,449 |
| **Total** | 1,175 | 3,758 | 4,933 |

**Results:**

There were 1,175 patients who had serum CRP above 3mg/L (“high” serum CRP) and 3,758 patients who had serum CRP of 3mg/L or below (“normal/low” serum CRP). The probability of dying within 4 years of study enrolment among patients with high serum CRP was 15.57%, which corresponds to an odds of dying within 4 years equal to 0.1845. This probability among patients with normal/low serum CRP was 8.01%, which corresponds to an odds equal to 0.0871. The odds ratio between the group with high serum CRP and the group with normal/low CRP is 2.1184, i.e. the odds of dying within 4 years among patients with high serum CRP tends to be 2.1184 times higher than such odds among patients with normal/low serum CRP. With 95% confidence, this result is consistent with a true odds ratio being anywhere between 1.7295 and 2.5898. Using a Fisher’s exact test, the result is statistically significant at the 0.05 critical threshold with a two-sided p-value of less than 0.0001. As a result, we reject the null hypothesis that the true odds ratio of dying within 4 years between the two defined groups are 1, in favor of an alternative hypothesis that the odds of dying within 4 years from study enrolment is associated with serum CRP level.

*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).*

**Methods:**

The association between serum CRP and all-cause mortality over the entire period of observation of these subjects were compared by looking at the entire survival distribution across the group with high serum CRP and the group with normal/low serum CRP. The difference in survival distributions was examined by performing a log-rank test. The log-rank test also examined difference in hazard functions across the two groups.

The instantaneous risk of death across these two groups were compared by investigating the hazard ratio (high serum CRP versus normal/low serum CRP). A 95% confidence interval was constructed using the Cox proportional hazards regression. The result was obtained directly from R using coxph command.

**Results:**

There were 1,175 patients who had serum CRP above 3mg/L (“high” serum CRP) and 3,758 patients who had serum CRP of 3mg/L or below (“normal/low” serum CRP). The survival distributions across the group with high serum CRP and the group with normal/serum CRP were illustrated in the following graph. Table 3 also depicts the Kaplan-Meier estimates of survival probabilities at important time points across the two groups.

From the graph as well as Table 3, we can see that the survival probabilities among patients with normal/low serum CRP were, at every time point, higher than the survival probabilities among patients with high serum CRP, as the survival curve for the normal/low serum CRP group was higher than and did not overlap over the analysis time with the survival curve for the high serum CRP group. A log-rank test suggests that the difference in the survival probabilities, hence followed by the difference in hazard, was statistically significant at the 0.05 critical threshold with a two-sided p-value less than 0.0001.

Using Cox proportional hazards regression, the instantaneous risk of death is estimated to be 68.7% greater for the group with high serum CRP compared to the group with normal/low serum CRP. With 95% confidence, this result was consistent with a true instantaneous risk of death being anywhere between 48.6% and 91.5% higher in the high serum CRP group compared to the normal/low serum CRP.

As a result, we reject the null hypothesis in favor of an alternative hypothesis that there is an association between serum CRP level and all-cause mortality.



**Table 3. Survival Probabilities at Selected Time Points**

|  |  |
| --- | --- |
|  | **Kaplan-Meier Estimates1** |
|  | **High Serum CRP** | **Normal/Low Serum CRP** |
| **Approximately 1 Year** | 0.966 (0.956 – 0.976) | 0.987 (0.984 – 0.991) |
| **Approximately 2 Years** | 0.926 (0.911 – 0.941) | 0.970 (0.965 – 0.976) |
| **Approximately 3 Years** | 0.881 (0.863 – 0.900) | 0.948 (0.941 – 0.955) |
| **Approximately 4 Years**  | 0.844 (0.824 – 0.865) | 0.920 (0.911 – 0.928) |
| **Approximately 5 Years** | 0.799 (0.776 – 0.822) | 0.884 (0.874 – 0.894) |
| **Approximately 6 Years** | 0.754 (0.729 – 0.780) | 0.852 (0.841 – 0.864) |

***1 Results presented are: point estimate (95% confident interval)***

*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:**

A priori I would have preferred in order to answer the question about an association between mortality and serum CRP is an analysis over the entire survival distributions across the three groups defined by the levels of serum CRP of the Mayo Clinics, due to the following reasons:

* It is much more precise not to have to dichotomize over survival times. Although in this case it is valid to dichotomize at 4-year time point, there are censored observations which can potentially have really long time to all-cause death. Note that the maximal follow-up time is roughly 8 years. There is a statistical, but not scientific in this case, to choose 4 years as a cut-off value of dichotomization.
* It is scientifically more pleasing and easier to understand, for people with scientific backgrounds, to condition on serum CRP level. Using 3 groups instead of 2 groups (high vs. normal/low) is scientifically more pleasing as well as giving a better balance of sample size.
* As the measurement of CRP value occurred before the recorded follow-up time, it is better to summarize the survival distribution.
* Summarizing the survival distributions also gives a better overall picture about what was going on throughout the entire analysis time. Dichotomizing observed time will only looking at the association between serum CRP and the vital status at only the chosen time point.