**Homework #1**

BIOST 518: Applied Biostatistics II

January 12, 2015

**Problem 1**

Given that the minimum time of follow-up among the censored observations is 1480 days, which is roughly just over four years, the vital status of each subject in the study is known at four years.

**Problem 2**

*Methods*: An indicator variable was created for death within four years of enrollment in the study (no subjects were censored during that period of observation). Descriptive statistics are described by groups based on serum C-reactive protein (CRP) levels (e.g., below 1 mg/L, between 1 – 3 mg/L, and above 3 mg/L), as well as in the overall sample. The mean, standard deviation, minimum and maximum are included within each group defined by serum CRP for continuous variables, such as age, BMI, and cholesterol. Percentages are used for binary variables, such as death within four years, gender, smoking history, and prior history of cardiovascular disease (CVD).

*Results*: Data is available on 5000 subjects. However, there are 67 subjects (including seven who died within four years) that have missing data with respect to serum CRP. Subjects with missing data are not included, or simply ignored at this point, in the analyses, though it is important to emphasize that the effect on the generalizability of the results due to these omissions cannot be evaluated. Out of the remaining 4933 subjects, six subjects were missing data on the variable, smoking history, three subjects were missing data on the variable, cholesterol, and 13 subjects were missing data on the variable, BMI.

From the 4933 subjects with available measurements, 428 had serum CRP levels less than 1 mg/L, 2629 had serum CRP levels between 1 and 3 mg/L, and 1876 had serum CRP levels greater than 3 mg/L. The table below illustrates the descriptive statistics within each group of serum CRP level. Subjects having serum CRP in the highest group (>3 mg/L) were least likely to be male than in other serum CRP levels. Cholesterol level was lowest among subjects with the lowest serum CRP levels, (205.9 mg/dL, compared to 212.5 mg/dL and 211.8 mg/dL, respectively, for 1-3 mg/L and >3 mg/L of serum CRP). The proportion of subjects with a smoking history was greatest among the highest serum CRP group (15.7%, compared to 9.6% and 10.0%, respectively, for <1 mg/L and 1-3 mg/L of serum CRP), and prior CVD were greatest among the highest serum CRP group as well (27.2%, compared to 18.2% and 20.6%, respectively, for <1 mg/L and 1-3 mg/L of serum CRP). Subjects with lower serum CRP levels appeared to have decreased mortality rate: 4.9% of subjects with a serum CRP level of less than 1 mg/L as compared to 13.9% of subjects with a serum CRP level of greater than 3 mg/L at the beginning of the study.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Serum C-Reactive Protein (CRP)** | | | | |
|  | **<1 mg/L**  **(n=428)** | **1-3 mg/L**  **(n=2629)** | **>3 mg/L**  **(n=1876)** | **Any Level**  **(n=4933)** |
| **Male (%)** | 45.5 | 44.5 | 37.7 | 42.0 |
| **Age (years)1** | 73.5 (5.80; 65 – 94) | 72.8 (5.57; 65 – 100) | 72.6 (5.50; 65 – 93) | 72.8 (5.56; 65 – 100) |
| **BMI (kg/m2) 1** | 23.8 (3.64; 15.6 – 38.6) | 26.1 (4.16; 14.7 – 53.2) | 28.1 (5.18; 15.3 – 58.8) | 26.7 (4.72; 14.7 – 58.8) |
| **Cholesterol (mg/dL) 1** | 205.9 (40.53; 109 – 407) | 212.5 (38.60; 73 – 363) | 211.8 (39.70; 96 – 430) | 211.7 (39.23; 73 – 430) |
| **Smoking history (%)** | 9.6 | 10.0 | 15.7 | 12.2 |
| **Prior CVD (%)** | 18.2 | 20.6 | 27.2 | 23.0 |
| **Death within 4 years (%)** | 4.9 | 7.7 | 13.9 | 9.8 |

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

*Methods*: An indicator variable was created for death within four years of enrollment in the study (no subjects were censored during that period of observation). Descriptive statistics are described by groups based on death within four years, survival for four years post-study entry, as well as in the overall sample. The mean, standard deviation, minimum and maximum are included within each group defined by vital status at four years for continuous variables, such as age, BMI, serum CRP, and cholesterol. Percentages are used for binary variables, such as gender, smoking history, and prior history of CVD.

*Results*: Data is available on 5000 subjects. However, there are 67 subjects (including seven who died within four years) that have missing data with respect to serum CRP. Subjects with missing data are not included, or simply ignored at this point, in the analyses, though it is important to emphasize that the effect on the generalizability of the results due to these omissions cannot be evaluated. Out of the remaining 4933 subjects, six subjects were missing data on the variable, smoking history, three subjects were missing data on the variable, cholesterol, and 13 subjects were missing data on the variable, BMI.

From the 4933 subjects with available measurements, 484 died within four years of the beginning of the study, while 4449 were still alive four years after the study began. The table below gives the descriptive statistics within each group. Subjects who died within four years post-study enrollment were more likely to be male, tended to be older, had lower cholesterol levels, and more likely to have a smoking history and prior CVD, as compared to subjects still alive at four years post-study enrollment. Subjects who died within four years also tended to have higher serum CRP levels at the start of the study: mean serum CRP was 5.4 mg/L among subjects who died within four years as compared to a mean serum CRP of 3.4 mg/L among subjects that were still alive after four years.

|  |  |  |  |
| --- | --- | --- | --- |
| **Vital Status at 4 Years Post-Study Enrollment** | | | |
|  | **Alive at 4 years**  **(n=4449)** | **Dead within 4 years**  **(n=484)** | **All Subjects**  **(n=4933)** |
| **Male (%)** | 39.9 | 60.4 | 41.9 |
| **Age (years)1** | 72.5 (5.32; 65 – 98) | 76.3 (6.71; 65 – 100) | 72.8 (5.60; 65 – 100) |
| **BMI (kg/m2) 1** | 26.7 (4.71; 14.7 – 58.8) | 26.3 (4.97; 14.8 – 48.1) | 26.7 (4.73; 14.7 – 58.8) |
| **Cholesterol (mg/dL) 1** | 212.5 (38.97; 78 – 430) | 204.1 (41.37; 73 – 396) | 211.7 (39.29; 73 – 430) |
| **Serum CRP (mg/L) 1** | 3.4 (5.87; 0 – 108) | 5.4 (8.10; 0 – 55) | 3.6 (6.15; 0 – 108) |
| **Smoking history (%)** | 11.8 | 14.3 | 12.1 |
| **Prior CVD (%)** | 21.0 | 41.4 | 23.0 |

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

**Problem 3**

*Methods*: Mean serum CRP levels were compared between subjects who died within four years of the start of the study and those who were still alive after four years. The t-test assuming unequal variances (i.e., Satterthwaite approximation) was used to test differences in the mean. 95% confidence intervals for the difference in population means were likewise based on that same assumption of unequal variances.

*Results*: Mean serum CRP was 3.42 mg/L among the 4449 subjects who were still alive after four years after study enrollment and 5.38 mg/L among the 484 subjects who died within four years. Based on a 95% confidence interval computed with an assumption of unequal variances, this observed result of 3.61 mg/L higher mean serum CRP among subjects dying earlier would not be unusual if the true difference population means were somewhere between a 3.44 mg/L to 3.78 mg/L higher mean serum CRP among subjects who die within four years. Using the t-test that assumes unequal variances, this result is statistically significant at a 0.05 level of significance (two-sided p-value < 0.0001), and we can with 95% confidence, reject the null hypothesis that the mean serum CRP levels are not different by vital status at four years post-enrollment in favor of a hypothesis that death within four years is associated with higher mean serum CRP.

**Problem 4**

*Methods*: Geometric mean serum CRP levels were compared between subjects who died within four years of the start of the study and those who survived at least four years. A t-test assuming unequal variance (i.e., Satterthwaite approximation) was used to test differences in the mean of log-transformed (base 10) serum CRP levels. 95% confidence intervals for the difference in population means for the log of serum CRP were likewise based on the same assumption of unequal variances. Given that some of the measurements of CRP were reported as zeroes, a correction of taking half of the value of the lower limit of detection for serum CRP was added to all zero values in order to calculate geometric means. Estimates and confidence intervals were then exponentiated in order to obtain inference on the geometric mean.

*Results*: Geometric mean serum CRP was 2.33 mg/L among the 4042 subjects who were still alive after four years after study enrollment and 3.22 mg/L among the 463 subjects who died within four years. Based on a 95% confidence interval computed with an assumption of unequal variances, this observed result of 27.5% lower geometric mean serum CRP among subjects surviving at least four years would not be unusual if the true ratio of population geometric means was somewhere between a 20.5% to 33.8% lower geometric mean serum CRP among subjects who are still alive after four years. Using the t-test that assumes unequal variances, this result is statistically significant at a 0.05 level of significance (two-sided p-value < 0.0001), and we can with 95% confidence, reject the null hypothesis that the geometric mean serum CRP levels are not different by vital status at four years post-enrollment in favor of a hypothesis that death within four years is association with higher geometric mean serum CRP.

**Problem 5**

*Methods*: The proportion of subjects dying within four years since the beginning of the study were compared between subjects who had serum CRP greater than 3 mg/L, and subjects whose serum CRP was less than or equal to 3 mg/L. Testing of differences in the probability of death within four years was done using Pearson’s chi-squared test for independence. Using Wald statistics, 95% confidence intervals for the difference in population four year mortality probabilities were calculated.

*Results*: From the 3057 subjects whose serum CRP levels were less than or equal to 3 mg/L, 7.29% were observed to die within four years, while 13.91% of the subjects with serum CRP levels greater than 3 mg/L died within four years after enrolling in the study. Based on a 95% confidence interval, this 6.62% higher absolute survival probability in subjects with a lower serum CRP level would not be unusual if the true difference in survival probabilities were between a 4.80% higher absolute probability of survival to a 8.44% higher absolute probability of survival in the low serum CRP group as compared to the high serum CRP group. Using a chi-squared test, this observation is statistically significant at a 0.05 level of confidence (two-sided p-value < 0.001), and we can with 95% confidence, reject the null hypothesis that the survival probabilities are not associated with serum CRP levels in favor of a hypothesis that survival probabilities are associated with serum CRP.

**Problem 6**

*Methods*: The odds of subjects dying within four years of being enrolled in the study were compared between subjects with serum CRP greater than 3 mg/L and subjects with serum CRP less than or equal to 3 mg/L. Testing of an odds ratio different from one was done using Fisher’s exact test. Exact methods were also used to compute 95% confidence intervals for the odds ratio.

*Results*: From the 3057 subjects with serum CRP less than or equal to 3 mg/L, the odds of dying within four years after the study began was 0.08 , while the remaining 1876 subjects with serum CRP greater than 3 mg/L, the odds of dying within four years post-enrollment was 0.16. Based on a 95% confidence interval, this observed odds ratio of 2.05 for the comparison of the high serum CRP group to the low serum CRP group would not be unusual if the true odds ratio was between 1.69 and 2.49. The Fisher’s exact test two-sided p-value of less than 0.0001 suggests that we can, with 95% confidence, reject the null hypothesis that the odds of four-year mortality are not associated with serum CRP in favor of a hypothesis that the odds of four-year mortality post-enrollment are association with serum CRP levels.

**Problem 7**

*Methods*: Using strata defined by serum CRP less than or equal to 3 mg/L and serum CRP greater than 3 mg/L, the survival distribution was estimated using Kaplan-Meier estimates. The logrank statistic was used to test the difference in survival distributions between those two groups. The hazard ratio and 95% confidence intervals were calculated using Cox proportional hazards regression with the Huber-White sandwich estimator of the standard errors.

*Results*: The graph and table below presents Kaplan-Meier estimates of survival probability for the 3057 subjects with serum CRP less than or equal to 3 mg/L and the 1876 subjects with serum CRP greater than 3 mg/L. The graph indicates a clear trend for higher survival probabilities for the lower serum CRP group at every point in time. The instantaneous risk of death is estimated to be 60.5% higher for the high serum CRP group compared to the low serum CRP group. Based on a 95% confidence interval, this observed hazard ratio of 1.60 comparing the high CRP group to the low CRP group would not be unusual if the true hazard ratio were between 1.43 and 1.80. A logrank test two-sided p-value of less than 0.0001 indicates that we can, with 95% confidence, reject the null hypothesis that the probability of survival is not associated with serum CRP levels in favor of a hypothesis that the probability of survival is associated with serum CRP levels.



|  |  |  |
| --- | --- | --- |
|  | **Survival Probabilities (Kaplan-Meier)** | |
|  | **CRP <3 mg/L** | **CRP >3 mg/L** |
| **1 year** | 0.9997 | 0.9995 |
| **2 years** | 0.9997 | 0.9989 |
| **3 years** | 0.9997 | 0.9973 |
| **4 years** | 0.9993 | 0.9973 |

**Problem 8**

Having done multiple models of analysis through this exercise, if I were to have selected a particular analysis in order to evaluate the association between 4-year mortality and serum CRP, I would firsdecide on what analysis I would do before I start any data collection or analysis in order to make sure that I am doing an *a priori* analysis. Insofar as what was done previously, my first point of consideration in doing an *a priori* analysis would be to keep the predictor of interest, serum CRP, as a continuous measurement, thereby preserving as much statistical precision as possible by not dichotomizing the variable. Second point to make is that when considering the association between 4-year mortality and serum CRP, temporality should matter. That is, the predictor of interest should be serum CRP and the outcome be mortality given that serum CRP measurements must occur earlier in time than death. From a scientific standpoint, it would be thus more logical to have serum CRP measurements precede death rather than to have death precede serum CRP measurements. Finally, in order to facilitate clear comprehension of the analysis, a comparison of the means and proportions would be easier to understand than using geometric means, odds ratios, or the hazard ratio. Taking all these considerations together, therefore, I would most likely answer the question about an association between mortality and serum CRP by comparing means of serum CRP across survival groups (i.e., survival/mortality at 4 years).