Homework 1
January 12, 2015

1. The minimum time to follow-up for data with censored observations is 1480 days. This is a little over 4 years, thus the vital status of each patient is known at four years making the proposed analysis valid.
2. **Methods**: Indicator variables were created for death within 4 years of the study enrollment, during which time no censored data were collected. Descriptive statistics are presented within groups defined by blood C reactive protein (CRP) levels (less than 1 mg/L, 1 and 3 mg/L inclusive, and greater than 3 mg/L) and the total sample population. Within each group defined by continuous variables, the mean, standard deviation, minimum and maximum are reported. For binary variables, percentages are reported.

**Results**: Data is available for 5000 participants, but only 4933 subjects had CRP data. These subjects were omitted from all of the analyses but it is unknown at this time how the omitted data affects the generalizability of the results. Of the 4933 subjects included in the analysis, it should also be noted that some were also missing data for BMI (13 subjects), smoking status (6 subjects) and cholesterol levels (3 subjects).

Within the study population, 428 had blood CRP levels below 1 mg/L, 2629 had blood CRP levels between 1-3 mg/L, and 1876 had blood CRP levels greater than 3 mg/L. The following table presents descriptive statistics within these groups as well as for the total population. There did not appear to be any trends between serum C reactive protein levels and age, sex, and cholesterol levels. In terms of BMI, smoking status, prior disease and death within 4 years, however, higher CRP levels appear to be associated with an increase in these variables. For example, 4.91% of the study subjects with less than 1 mg/L CRP died within 4 years compared to 13.91 % of study participants with greater than 3 mg/L CRP.

|  |  |
| --- | --- |
|  | **Serum C Reactive Protein (CRP)** |
|  | **< 1 mg/L(n=428)** | **1-3 mg/L****(n=2629)** | **< 3 mg/L****(n=1876)** | **Any****(n=4933)** |
| **Age (yrs)1** | 73.45 (5.80; 65-94) | 72.81 (5.57; 65-100) | 72.63 (5.50; 65-93) | 72.80 (5.56; 65-100) |
| **Male (%)** | 45.6 % | 44.5 % | 37.7 % | 42.0 % |
| **BMI1** | 23.82 (3.64; 15.6-38.6) | 26.08 (4.16; 14.7-53.2) | 28.11 (5.18; 15.3-58.8) | 26.66 (4.72; 14.7-58.8) |
| **Smoker (%)** | 9.6 % | 10.1 % | 15.73 % | 12.18 % |
| **Cholesterol (mg/L)1** | 205.99 (40.53; 109-407) | 212.49 (38.6; 73-363) | 211.84 (39.4; 96-430) | 211.68 (39.23; 73-430) |
| **Prior Disease (%)**  | 18.22 % | 20.62 % | 27.24 % | 22.93 % |
| **Death within 4 years** | 4.91 % | 7.68 % | 13.91 % | 9.81 % |

1Continuous variables are reported as the mean (standard deviation; minimum-maximum)

1. **Methods**: Mean CRP levels were compared between subjects who died within 4 years of enrollment and those who lived for at least 4 years post-enrollment. Differences in the mean were tested using a t test which allowed for unequal variances. 95% confidence intervals for the difference in population means were also calculated with allowances for unequal variances between the two populations.

**Results**: Mean CRP was 3.42 mg/L among 4449 subjects who survived for at least 4 years after enrollment while mean CRP was 5.38 mg/L among the 484 subjects who died within 4 years. The observed tendency for subjects who died earlier to have 1.95 mg/L higher mean CRP than those who survived would not be unusual if the true difference in population CRP means were between 1.21 mg/L and 2.70 mg/L, based on 95% confidence intervals calculated while allowing for unequal variances. Similarly, the results of t test that allowed for the possibility of unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P=0). This allows us to reject the null hypothesis that the mean CRP levels are not different by vital status at 4 years in favor of a hypothesis that death within 4 years is associated with higher mean CRP levels.
2. **Methods**: Geometric mean CRP levels were compared between subjects who died within 4 years of enrollment and those who lived for at least 4 years post-enrollment. The CRP levels were log-transformed and it should be noted that log-transforming the data generated 496 missing values as the CRP values reported were 0 (407 subjects who survived for 4 years and 21 subjects who died within 4 years). These subjects were excluded from this analysis. Differences in the mean of log transformed CRP levels were tested using a t test which allowed for unequal variances. 95% confidence intervals for the difference in population means were also calculated with allowances for unequal variances between the two populations. Finally, estimates and confidence intervals were exponentiated to obtain inferences on the geometric mean.

**Results**: Geometric mean CRP was 2.34 mg/L among 4042 subjects who survived for at least 4 years after enrollment while the geometric mean CRP was 3.22 mg/L among the 463 subjects who died within 4 years. Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 27.44% lower geometric mean among subjects who survived at least 4 years would not be judged unusual if the true ratio of population geometric means indicated anywhere between a 20.47% to 33.81% lower geometric mean CRP among subjects who survived for at least 4 years. Using a t test on log transformed CRP that similarly allows for the possibility of unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0000), and we can with high confidence reject the null hypothesis that the geometric mean serum CRP levels are not different by vital status at 5 years in favor of a hypothesis that death within 4 years is associated with higher geometric mean CRP levels.
3. **Methods**: The proportion of subjects dying within 4 years of study enrollment were compared between subjects who had serum CRP greater than 3 mg/L and subjects whose serum CRP was measured to be less than 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 the difference in population 5 year mortality probabilities were computed using Wald statistics.

**Results**: Of the 1876 study subjects who had serum CRP levels at or above 3 mg/L, 13.9% were observed to die within 4 years after study enrollment while 7.3% of the subjects who had serum CRP levels below 3 mg/L were observed to die within 4 years. Based on a 95% confidence interval, the 6.6 % observed difference in mortality probability would not be judged as unusual if the true difference in mortality probabilities were 4.8% to 8.4% higher among subjects with CRP levels greater than or equal to 3 mg/L. Using a chi squared test, this observations was statistically significant at a 0.05 level of significance (P=0.0000), and we can reject the null hypothesis that the survival probabilities are not associated with CRP levels with high confidence.
4. **Methods**: The odds of subjects drying within 4 years of study enrollment were compared between subjects who had serum CRP levels at or above 3 mg/L and those who had less than 3 mg/L. An odds ratio different from 1 was tested using Fisher’s exact test and 95% confidence intervals for the odds ratio was also computed using exact methods.

**Results**: The odds ration obtained using Fisher’s exact test was 2.05. Based on 95% confidence intervals, the observed odds ratio of 2.05 for the comparison of the high CRP group with the low CRP group would not be be unusual if the true odds ratio was between 1.69 and 2.49. A Fisher’s exact test two-sided p-value of 0.0000 suggests that we can reject the null hypothesis that the odds of 4 year mortality are not associated with serum CRP levels with high confidence.
5. **Methods**: The survival distribution was estimated using Kaplan-Meier estimates, stratified by serum CRP less than 3 mg/L (low) or greater than and equal to 3 mg/L (high). Difference in survival distributions between the two groups were tested using the logrank statistics. The hazard ratio and 95% confidence intervals was computed using Cox proportional hazards regression.

**Results:** The following graph and table depict Kaplan-Meier estimates of survival probability for 1876 participants with serum CRP levels greater than/equal to 3 mg/L and 3057 individuals with serum CRP levels less than 3 mg/L. From the graph and the table, it is apparent that there is a tendency for individuals who have low serum CRP levels to higher survival probabilities, both within the first 4 years following enrollment as well as subsequent years. The instantaneous risk of death is estimated to be 22.11% higher for the high CRP group compared to the low CRP group. Based on a 95% confidence interval, this observed hazard ratio of 1.2211 when comparing the high CRP group to the low CRP group would not be judged unusual if the true hazard ratio were anywhere between 1.15 to 1.29. A p value of 0.0000 obtained for the logrank two-sided test suggests that with high confidence, we can reject the null hypothesis that the probability of survival is not associated with serum CRP levels.



|  |  |
| --- | --- |
|  | Kaplan-Meier Survival Probabilities |
|  | CRP < 3 mg/L | CRP ≥ 3 mg/L |
| 1 year | 0.9892 | 0.9712 |
| 2 years | 0.9751 | 0.9355 |
| 3 years | 0.9532 | 0.8977 |
| 4 years | 0.9271 | 0.8609 |

1. *A priori*, I would have said that comparing mean serum CRP levels across survival groups would be the first analysis that comes to mind. It is likely that comparing the geometric means may be more precise than comparing simple means, however it is also more technical and requires more careful interpretation of the results due to it logarithmic nature. Comparing survival estimates between groups with low and high serum CRP levels would yield some useful results as it would become quickly apparent (just from looking at the Kaplan-Meier curve) that serum CRP levels are associated with survival. The problem with this, however, is that it requires that a continuous variable (i.e. serum CRP levels) is broken down into two categories, which results in some loss of precision.