**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

**Homework #1**

Due: 1/13/2013

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 5 years of study enrolment or death after 5 years. Why is this valid? Provide descriptive statistics that support your answer.

**Answer:** It is valid to dichotomize the time to death according to death within 5 years or after 5 years because the time to first censoring was greater than 5 years (the first censoring event was at 1827 days). Therefore we can divide our population into those that died within five years and those that did not.

Table 1 provides the counts of those that died and those that were censored within 5 years or after 5 years.

**Table 1.**

|  |  |  |
| --- | --- | --- |
|  | Number dead | Number censored |
| Within 5 years | 121 | 0 |
| Over 5 years | 12 | 602 |

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 LDL and 5 year all-cause mortality in the medical literature. In attention to the two variables of primary interest, you may restrict attention to age, sex, weight, smoking history, and prior history of cardiovascular disease (coronary heart disease (CHD), congestive heart failure (CHF), and stroke.

**Answer:**

*Methods:* Descriptive statistics including the mean, standard deviation, and proportion of events were computed for time to death (within or after 5 years), age, sex, weight, smoking history, and history of congestive heart failure, coronary heart disease, and stroke stratified by subjects low density lipoprotein (LDL) values (less than or greater than 160 mg/dl). This was selected as the cutoff because any serum LDL value greater than 160mg/dl is considered high by the Mayo Clinic. Note that 10 out of the 735 participants have missing LDL values and thus were not used for this analysis or further analyses.

*Inference:* Table 2 contains various descriptive statistics. It appears that many of the values are similar between the two groups except perhaps sex and history of congestive heart failure, however no t-tests were run to determine if these differences are significant. We also note that the number of subjects with high LDL is much lower than those with low to moderate LDL, which could lead to volatility in our results.

**Table 2**

|  |  |  |
| --- | --- | --- |
| Characteristics | **Low-Moderate LDL (N=618)****(<160mg/dl)**Mean ± SD | **High LDL (N=107)****(>160mg/dl)**Mean ± SD |
| Time to Death within 5 years after 5 years | 0.170.83 | 0.130.87 |
| Age (years) | 74.51 ± 5.39 | 74.88 ± 0.42 |
| Sex male female | 0.510.49 | 0.420.58 |
| Weight (lbs) | 159.36 ±30.78 | 162.74 ± 30.68 |
| Congestive heart failure no history history | 0.940.06 | 0.970.03 |
| Smoking Status never smoker former smoker current smoker | 0.430.430.14 | 0.460.420.12 |
| Coronary heart disease none angina MI | 0.790.090.12 | 0.800.070.12 |
| Stroke none transient ischemic attack stroke | 0.880.030.10 | 0.810.060.13 |

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years.

**Answer:**

*Methods:* A two-sample t-test that does not assume equal variances, comparing the mean values of LDL in groups defined by whether or not the subject died within 5 years of the study was performed. The null hypothesis is that the mean LDL of those that died within 5 years(1) minus the mean LDL of those that died after 5 years(2) is 0. The alternative hypothesis is that the mean LDL of those that died within 5 years(1) minus the mean LDL of those that died after 5 years(2) is nonzero.

H0: μ1 – μ2 = 0

Ha: μ1 – μ2 ≠ 0

*Inference:* Mean LDL for those that died within 5 years is 118.70 mg/dl (95% CI: 112.13, 125.26; n=119) and mean LDL for those that died after 5 years is 127.20 mg/dl (95% CI: 124.57, 129.83; n=606). The mean difference is -8.50 mg/dl (95% CI: -15.56, -1.44) with a corresponding p-value is 0.0186. This is statistically significant at an alpha level of 0.05. Therefore, we reject the null hypothesis that mean LDL are the same for those that died within 5 years and those that died after 5 years. We have sufficient evidence that the mean LDL is different between these two groups. Our results are consistent with a mean difference in LDL of those that died within 5 years is between 1.44 mg/dl to 15.56 mg/dl lower than those that died after 5 years.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing geometric mean LDL values across groups defined by vital status at 5 years.

**Answer:**

*Methods:* A two-sample t-test that does not assume equal variances, comparing the ratio of geometric mean values of LDL in subjects that died within 5 years and those that died after 5 years was performed. The null hypothesis is that the ratio of the geometric mean LDL of those that died within 5 years(1) and the geometric mean LDL of those that died after 5 years(2) is 1. The alternative hypothesis is that the ratio of geometric mean LDL of those that died within 5 years(1) and the geometric mean LDL of those that died after 5 years(2) is not one.

H0: μ1/μ2 = 1

Ha: μ1/μ2 ≠ 1

*Inference:* The geometric mean LDL for those that died within 5 years is 112.01 mg/dl (95% CI: 104.54, 120.02; n=119) and geometric mean LDL for those that died after 5 years is 122.83 mg/dl (95% CI: 120.21, 125.50; n=606). The geometric mean ratio is 0.91 (95% CI: 0.85, 0.98) with a corresponding p-value of 0.0128. This is statistically significant at an alpha level of 0.05. We reject the null hypothesis that geometric mean LDL are the same for those that died within 5 years and those that died after 5 years. We have sufficient evidence that the geometric mean LDL is different between these two groups.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL).

**Answer:**

*Methods:* A chi-squared test was performed to determine whether dichotomized LDL values and death within 5 years are independent. Furthermore, a binomial Wald confidence interval was calculated for the risk difference. Here we note that a chi-squared test and z-test for two binomial proportions are equivalent. Our null hypothesis is that dichotomized LDL values are independent of time to death. The alternative hypothesis is that dichotomized LDL values and time to death are not independent.

H0: LDL values greater than or equal to 160 mg/dl and LDL values less than 160 mg/dl are independent of death within and over 5 years.

Ha: LDL values greater than or equal to 160 mg/dl and LDL values less than 160 mg/dl are not independent of death within and over 5 years and time to death.

*Inference:* Table 5 displays the corresponding contingency table. The proportion of subjects with LDL values less than 160 mg/dl that die within 5 years is 105/618 or 0.17. The proportion of subjects with LDL values greater than or equal to 160 mg/dl that die within 5 years is 14/107 or 0.13. The risk difference is 0.04 (95% CI: -0.03,0.11) and the corresponding two-sided p-value is 0.314. This is not less than an alpha level of 0.05; therefore, we fail to reject the null hypothesis that dichotomized LDL values and probability of death within 5 years is independent. Our results are consistent with a true risk difference of dying within 5 years is between 0.03 lower and 0.11 higher in subjects with LDL values less than 160 mg/dl.

**Table 5.**

|  |  |  |
| --- | --- | --- |
|  | LDL < 160 mg/dl | LDL > 160 mg/dl |
| Death within 5 years | 105 | 14 |
| Death over 5 years | 513 | 93 |

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL).

**Answer:**

*Methods:* The odds ratio (OR) was computed and corresponding confidence intervals were calculated using the methods described by Woolf. The p-value was calculated in the same manner as question 5 (using the chi-squared test). The null hypothesis is that the odds ratio is one. The alternative hypothesis is that the odds ratio is not one.

H0: OR = 1

Ha: OR ≠ 1

*Inference:*

The odds ratio of death within 5 years is 1.36 (95% CI: 0.75, 2.50). The odds of death within 5 years is 36% higher for the group with LDL values less than 160 mg/dl. With 95% confidence, the observed results would not be unusual if the true odds of death within 5 years for those with low LDL was between 25% lower to 150% higher compared to those with higher LDL values. The calculated two-sided p-value is 0.314. Since this is not less than an alpha level of 0.05 we fail to reject the null hypothesis that the odds ratio is 1.

1. Perform a statistical analysis evaluating an association between serum LDL 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 LDL (“high” = LDL > 160 mg/dL).

**Answer:**

*Methods:* A log-rank test for equal hazard curves was performed. This is equivalent to a log-rank test for equal survival curves. The null hypothesis is that the hazard curve for subjects with LDL less than 160 mg/dl (1) and the hazard curve for subjects with LDL greater than or equal to 160 mg/dl (2) are equivalent for all time. The alternative hypothesis is that the hazard curves are not the same for all time.

H0: h1(t) = h2(t) for all t

Ha: h2(t) ≠ h2(t) for all t

**Figure 7**

****

*Inference:* The smoothed instantaneous hazard estimates, including relevant confidence intervals are shown in Figure 7. Table 7 displays survival estimates. Note that the survival estimates do not go beyond 0.25, thus we cannot estimate a median survival time. Based on the two-sided p-value of 0.22, we fail to reject the null hypothesis that the hazard curves for those with LDL less than 160 mg/dl and those with LDL greater than or equal to 160 mg/dl are the same. We do not have sufficient evidence that the instantaneous risk of death for those with high LDL is different from those with lower LDL.

**Table 7**

|  |  |  |
| --- | --- | --- |
| Time (years) | **Low-Moderate LDL (N=618)****(<160mg/dl)**Survival Estimate (95% CI) | **High LDL (N=107)****(>160mg/dl)**Survival Estimate (95% CI) |
| 1 | 0.981 (0.966, 0.989) | 1 |
| 2 | 0.952 (0.931, 0.966) | 0.981 (0.927, 0.995) |
| 3 | 0.918 (0.893, 0.937) | 0.953 (0.891, 0.980) |
| 4 | 0.887 (0.859, 0.909) | 0.907 (0.833, 0.949) |
| 5 | 0.830 (0.798, 0.858) | 0.869 (0.789, 0.920) |

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 LDL? Why?

**Answer:**

Both LDL and time to death are continuous, and we therefore lose information when we dichotomize these variables. Therefore, I would opt to use some sort of regression. Since we are working with censored data the proportional hazards regression would be a good choice. There are a number of benefits to choosing regression, namely that we will not have to dichotomize our data, and we can also easily control for possible confounders.

In this assignment, questions 3 and 4 asked us to dichotomize our data by death within 5 years. Questions 5 and 6 asked us to dichotomize our data by both low and high levels of LDL and time to death. Question 7 asked us to dichotomize our data by LDL. If I were to select one of the methods we have covered in class, I would select comparing the mean LDL values between those that died within 5 years and those that died after 5 years. This method will require that only time to death be dichotomized. Given that over 600 of the subjects were right-censored after 5 years, this seems like a reasonable dichotomization. The geometric mean does offer some advantages since it is less influenced by outliers, but the ratio of geometric means can be difficult to interpret.