Biostat 515 Homework 1

1. Although the observations of time to death are subject to right censoring, we can dichotomize the time to death as a binary variable since our questions are not interested in time to death, but rather the 5 year mortality rate. This is possible since there are no censoring events before 5 years (or 1826.25 days according to the variable obstime which measures the observation time in days if we consider each year to be 365.25 days), and we assume non-informative censoring. As shown in the table below, there are no deaths (and hence no censoring) for subjects whose obstime is under 5 years.

**Death status dichotomized by 5 years of obstime**

|  |  |  |  |
| --- | --- | --- | --- |
|   | **death** |   |   |
| **obstime under 5 yrs** | 0 | 1 | Total |
| 0 | 602 | 12 | 614 |
| 1 | **0** | 121 | 121 |
| Total | 602 | 133 | 735 |

1. The data set includes information on 735 subjects, 614 of which die before year 5 and 121 of which survive past year 5. The number of missing data doesn’t seem to be substantial as there are just 10 total missing values for ldl and 1 missing data for packyrs. The mean and SD were rounded to two decimal places for better display.

We are interested in exploring the association between serum LDL and 5 year all-cause mortality, as well as between serum LDL and other pertinent variables (shown below). The mean ldl level for subjects who die before year 5 is 127.20 with a SD of 32.93, and the mean ldl level for subjects who survive past year 5 is 118.70 with a SD of 36.16. Considering the high standard deviation and the documentation which reports 100 to 189 mg/dL as typical measures of LDL for persons over age 70 (typical in our sample), it is hard to determine association.

Especially if we take other pertinent variables (potential confounders) into account, we discover that variables packyrs, chd, chf, and stroke (smoking history in pack years, history of coronary heart disease, congestive heart failure, and stroke, respectively) are all associated with death before year 5, casting doubts on the prospect that LDL levels and 5 year all-cause mortality exhibit significant associations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1: Descriptive Statistics for selected variables** |  |  |  |
|   | **variables** | **N** | **mean** | **sd** | **iqr** | **min** | **max** |
| **death before year 5** | **ldl** | 606 | 127.20 | 32.93 | 45 | 39 | 247 |
|  | **age** | 614 | 74.19 | 5.22 | 6 | 65 | 99 |
|   | **male** | 614 | 0.47 |  |  |  |  |
|   | **weight** | 614 | 160.11 | 30.35 | 41.50 | 74 | 258 |
|   | **packyrs** | 614 | 17.95 | 24.69 | 31.88 | 0 | 180 |
|   | **chd** | 614 | 0.28 | 0.64 |  | 0 | 2 |
|   | **chf** | 614 | 0.04 |  |  |  |  |
|   | **stroke** | 614 | 0.18 | 0.55 |  | 0 | 2 |
| **death after year 5** | **ldl** | 119 | 118.70 | 36.16 | 46.00 | 11 | 227 |
|   | **age** | 121 | 76.48 | 6.17 | 9.00 | 67 | 91 |
|   | **male** | 121 | 0.64 |  |  |  |  |
|   | **weight** | 121 | 159.12 | 32.79 | 37.00 | 96 | 264 |
|   | **packyrs** | 120 | 28.05 | 36.04 | 46.00 | 0 | 240 |
|   | **chd** | 121 | 0.62 | 0.85 |  | 0 | 2 |
|   | **chf** | 121 | 0.14 |  |  |  |  |
|   | **stroke** | 121 | 0.52 | 0.85 |  | 0 | 2 |
| **Total** | **ldl** | 725 | 125.80 | 33.60 | 45.00 | 11 | 247 |
|   | **age** | 735 | 74.57 | 5.45 | 7.00 | 65 | 99 |
|   | **male** | 735 | 0.50 |  |  |  |  |
|   | **weight** | 735 | 159.95 | 30.74 | 40.50 | 74 | 264 |
|   | **packyrs** | 734 | 19.60 | 27.11 | 33.75 | 0 | 240 |
|   | **chd** | 735 | 0.33 | 0.69 |  | 0 | 2 |
|   | **chf** | 735 | 0.06 |  |  |  |  |
|   | **stroke** | 735 | 0.24 | 0.62 |  | 0 | 2 |

1. We use a standard t-test with the assumption of unequal variances to compare the mean LDL levels by death status after 5 years. We dichotomize the variable obstime into before 1826.25 days (5 years) and after 1826.25 days using the variable deadin5 in order to stratify death status after 5 years, which is valid since the first censoring event in variable obstime is not until time 1827. LDL level is a continuous variable and death status is a binary variable, so it makes sense to use the t-test. Since the problem did not specify whether we should assume equal or unequal variances across the two groups, we assume the latter.

|  |  |  |
| --- | --- | --- |
| **Two-sample t-test with unequal variances** |  |  |
| **Group** | **Obs** | **Mean** | **Std. Err.** | **Std. Dev.** | **[95% Conf.** | **Interval]** |
| 0 | 606 | 127.198 | 1.338 | 32.929 | 124.571 | 129.825 |
| 1 | 119 | 118.698 | 3.315 | 36.157 | 112.134 | 125.261 |
| **combined** | 725 | 125.803 | 1.248 | 33.602 | 123.353 | 128.253 |
| **diff** |   | 8.501 | 3.574 |   | 1.441 | 15.560 |
| **p-value = 0.0186** |  |  |  |  |  |

The mean LDL level is estimated to be 127.20 mg/dL among subjects who survive at least 5 years, and 118.70 mg/dL among subjects who die within 5 years. Comparing the two groups, we can estimate that the mean LDL level is 8.50 mg/dL higher among subjects who survive at least 5 years relative to those who die within 5 years. This observed difference is statistically different from 0 (P=0.0186), with a 95% confidence interval suggesting that the observed difference is beyond some random coincidence if the true difference was between 1.44 mg/dL and 15.56 mg/dL, with the survivors averaging higher levels of LDL.

1. We use a standard t-test with the assumption of unequal variances on log-transformed data to compare the geometric mean LDL levels by death status after 5 years. We generate a variable logldl which is the log-transformation of variable ldl, then back-transform our estimates to get the final result. Since the problem did not specify whether we should assume equal or unequal variances across the two groups, we assume the latter.

|  |  |  |
| --- | --- | --- |
| **Two-sample t-test with unequal variances** |  |  |
| **Group** | **Obs** | **Mean** | **Std. Err.** | **Std. Dev.** | **[95% Conf.** | **Interval]** |
| 0 | 606 | 4.811 | 0.011 | 0.270 | 4.789 | 4.832 |
| 1 | 119 | 4.719 | 0.035 | 0.380 | 4.650 | 4.788 |
| **combined** | 725 | 4.796 | 0.011 | 0.293 | 4.774 | 4.817 |
| **diff** |   | 0.092 | 0.037 |   | 0.020 | 0.164 |
| **p-value = 0.0128** |  |  |  |  |  |

The geometric mean LDL level is estimated to be 112.01 mg/dL among subjects who die before 5 years and 122.83 mg/dL among subjects who survive past 5 years. Comparing the two groups, we estimate that the geometric mean cholesterol is 9.65% higher among subjects who survive past 5 years relative to subjects who die before 5 years. This observed difference is statistically different from 0 (P=0.0128), with a 95% confidence interval suggesting that the observed difference is beyond some random coincidence if the true geometric mean LDL of survivors was between 2.01% and 17.87% higher than that for nonsurvivors. In this case, we reject the null hypothesis of no association between LDL levels and death status in favor of a trend toward higher geometric mean LDL among subjects who survive past year 5.

1. We can use the chi square test to test for the risk difference (or mortality) between the group of subjects with high serum LDL (LDL > 160 mg/dL) and the group of subjects without high serum LDL (LDL < 160 mg/dL). The chi squared test assumes the independence of two samples, and thus our null hypothesis is that the true absolute difference in risk is 0% between the two groups.

14 of 107 patients (13.1%) with high serum LDL (LDL > 160 mg/dL) died within 5 years, while 105 of 618 patients (17.0%) without high serum LDL died within 5 years. Based on the chi squared test, the p-value of 0.3139 suggests that the observed absolute difference of 3.9% is not enough to reject the null hypothesis that the true absolute difference in risk is 0% between the two groups. The 95% CI for difference in mortality rates (mortality rate for patients with high serum LDL minus mortality rate for patients without high serum LDL) is between -10.95% and 3.14%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ldl > 160** | **ldl < 160** | **Total** |
| **Cases** | 14 | 105 | 119 |
| **Noncases** | 93 | 513 | 606 |
| **Total** | 107 | 618 | 725 |
| **Risk** | 0.131 | 0.170 | 0.164 |
| **Point estimate of risk difference = -0.039** |
| **95% CI = (-0.1095, 0.0314)** |  |

1. Since we want to compare the distinct odds of survival for each group of subjects (with high serum LDL and non-high serum LDL), we use the Stata command cc which computes point estimates and confidence intervals for the odds ratio along with the value of chi square test.

When comparing the group of subjects with high serum LDL (LDL > 160 mg/dL) and the group of subjects without high serum LDL (LDL < 160 mg/dL), the odds of dying within 5 years is estimated to be 26.45% lower (odds ratio 0.7355) for subjects with high serum LDL. We cannot conjecture that the observed difference is statistically different from an odds ratio of 1 (the null hypothesis) due to the p-value of 0.3139, with the 95% CI suggesting that the observed odds ratio is what we typically expect to observe if the true odds ratio was between 62.7% lower and 36.1% higher for the group with high serum LDL.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **cc test** |  |  |  |  |
|   |   |   |   | **Proportion** |
|   | **Exposed** | **Unexposed** | **Total** | **Exposed** |
| **Cases** | 14 | 105 | 119 | 0.118 |
| **Controls** | 93 | 513 | 606 | 0.154 |
| **Total** | 107 | 618 | 725 | 0.148 |
| **odds ratio = 0.7355** |  | **95% CI = (0.373, 1.361)** |  |

1. Since we want to compare the distinct instantaneous risk of death for the group of subjects with high serum LDL and the group of subjects without high serum LDL, we can use the Logrank test for the equality of survivor functions on the binary variable highldl.

The null hypothesis of the Logrank test is that the group of subjects with high serum LDL (subjects with LDL > 160 mg/dL) and the group of subjects without high serum LDL (subjects with LDL < 160 mg/dL) have identical survivor functions, or equal instantaneous risk of death at every timepoint. Based on the p-value of 0.2249 we cannot reject the null hypothesis that the two groups have the same survivor functions. We would proceed our study based on hypothesis favoring the insignificance of high serum LDL on death.

|  |
| --- |
| **Logrank test for equality of survivor functions** |
|   | **Events** | **Events** |  |
| **highldl** | **observed** | **expected** |  |
| **0** | 116 | 111.01 |  |
| **1** | 15 | 19.99 |  |
| **Total** | 131 | 131 |  |
| **chi2 value = 1.47** |  |  |
| **p-value = 0.2249** |  |  |

1. Personally, I would have preferred the logrank test. Preferably, we want to condition on past data (LDL levels) and predict the mortality rate after 5 years. The t-tests we performed in #3 and #4 condition on death status, which is not ideal although they lead to significant p-values. The chi-squared test to examine the probability and odds of death based on LDL status (#5 and #6) seem like better options, as we can condition on past data and predict future results. However, one important consideration is that the data is right-censored, and the logrank test allows us to adjust for censoring.