**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #1**

January 6, 2014

**Written problems:** To be submitted as a MS-Word compatible email attachment to semerson@uw.edu by 9:30 am on Monday, January 13, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

***In all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***
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.*

This is valid because we have not lost any patients to follow-up. There are 735 total patients, all of whom are accounted for in our data set when we dichotomize time to death to 5 years within study enrollment or death after 5 years. One hundred and twenty-one passed away before 5 years, and 614 passed away after 5 years for a total of 735 patients.

|  |  |  |
| --- | --- | --- |
| Survival (>5 years) | Frequency | Percent |
| No | 121 | 16.46 |
| Yes | 614 | 83.54 |
| Total | 735 | 100 |

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.*

Time to death was dichotomized into less than 5 years and more than five years of observation time. History of coronary heart disease (CHD) was also collapsed into two groups, those with no history and those with either a history of angina or coronary disease. Also, stroke was dichotomized into either no history of a history of a transient ischemic attack or stroke.

A total of 121 patients did not survive beyond 5 years. This group had a higher proportion of male patients, 64.5% vs. 46.9% and smoked more cigarettes with an average number of pack years smoked of 28 vs. 17.9. The group that did not survive beyond 5 years also had higher rates of cardiovascular disease. The proportion with CHD was 38% vs. 18% in the group that survived beyond 5 years, and 14% with CHF vs. 4%, and 29% with a history of stroke compared with 10.4% in the other group.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Frequency | Mean | SD | Minimum | Median | Maximum |
| **Survival < 5 years** |  |  |  |  |  |  |
| LDL (in mg/dL) | 119 | 118.7 | 36.2 | 11 | 117 | 227 |
| Age (in years) | 121 | 76.5 | 6.2 | 67 | 75 | 91 |
| % Male | 121 | 64.5 | 0.5 | 0 | 1 | 1 |
| Weight (in lbs) | 121 | 159.1 | 32.8 | 96 | 154 | 264 |
| Number of pack years smoked | 120 | 28.0 | 36.0 | 0 | 18.4 | 240 |
| % History of CHD | 121 | 38.0 | 0.5 | 0 | 0 | 1 |
| % History of CHF | 121 | 14.0 | 0.3 | 0 | 0 | 1 |
| % History of stroke | 121 | 28.9 | 0.5 | 0 | 0 | 1 |
|  |  |  |  |  |  |  |
| **Survival > 5 years** |  |  |  |  |  |  |
| LDL (in mg/dL) | 606 | 127.2 | 32.9 | 39 | 127 | 247 |
| Age (in years) | 614 | 74.2 | 5.2 | 65 | 73 | 99 |
| % Male | 614 | 46.9 | 0.5 | 0 | 0 | 1 |
| Weight (in lbs) | 614 | 160.1 | 30.3 | 74 | 158.9 | 258 |
| Number of pack years smoked | 614 | 17.9 | 24.7 | 0 | 4.4 | 180 |
| % History of CHD | 614 | 17.8 | 0.4 | 0 | 0 | 1 |
| % History of CHF | 614 | 3.9 | 0.2 | 0 | 0 | 1 |
| % History of stroke | 614 | 10.4 | 0.3 | 0 | 0 | 1 |

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.*

I used a two-sample two-sided t-test with equal variance where the null hypothesis stated there was no difference between the mean LDL values across the two groups by survival at 5 years. The alternative hypothesis was that there was a difference in mean LDL between the two survival groups.

We have evidence to reject the null hypothesis and find a statistically significant association between mean LDL and 5-year all-cause mortality (p=0.0115). Patients who survived < 5 years from the start of the study had a mean LDL that was 8.5mg/dL (with a 95% confidence interval of 1.9 to 15.1) LESS than the group that had survived beyond 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.*

A two-sample two-sided t-test with unequal variance was performed to compare the geometric mean LDL across groups defined by time to death within 5 years. A log value of LDL was generated and the t-test was run. The point value and confidence interval were then exponentiated to estimate the geometric mean. The null hypothesis is that there is no difference in the geometric mean LDL across the two survival groups. The alternative hypothesis is that there is a difference in the geometric mean LDL across the two survival groups.

There is evidence to reject the null hypothesis. There is a statistically significant difference in the geometric mean of 0.0002 with a 95% confidence interval of <0.0001 to 0.2366 (p = 0.013). That is, patients who survived < 5 years had a geometric mean LDL that was 0.0002 LESS than the group that survived beyond 5 years.

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).*

LDL was dichotomized into groups where LDL was > 160mg/dL and < 160 and compared to the dichotomized survival variable. The confidence intervals can be calculated using the binomial option. The probability of survival within 5 years when LDL is <160 mg/dL is 83.0% with a 95% confidence interval of 79.8% to 85.9%. The probability of survival within 5 years when LDL is > 160 mg/dL is 86.3% with a 95% confidence interval of 78.7% to 92.0%.

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).*

LDL was dichotomized into groups where LDL was > 160mg/dL and < 160 and compared to the dichotomized survival variable. As these are two binary variables, we can use the command for cohort studies or logistic regression. We calculated that the odds of survival at 5 years is 1.29 times higher (95% confidence interval 0.732 to 2.279) for patients with a serum LDL above 160mg/dL (p=0.375). In other words, the odds of surviving at 5 years is estimated to be 29% higher when LDL is greater than or equal to 160 mg/dL. The group with the lower LDL tends toward poorer survival at 5 years. These differences are not statistically significant from an odds ratio of 1. We have insufficient evidence to reject the null hypothesis that there is no association between survival time and LDL levels.

|  |  |  |  |
| --- | --- | --- | --- |
|  | LDL>160 | LDL <160 | Total |
| Survival < 5 years | 105 | 16 | 121 |
| Survival >5 years | 513 | 101 | 614 |
| Total | 618 | 117 | 735 |

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).*

The instantaneous risk of death is best measured using a proportional hazards regression model of the censored time to death on untransformed LDL levels with robust standard error estimates. When comparing the group with the “high” LDL with the normal to low LDL, the instantaneous risk of death is estimated to be 25.0% lower (hazard ratio 0.750) in the “high” LDL group. This difference is statistically different from a hazard ratio of 1 with a 95% confidence interval of 0.450 to 1.251, where the group with the LDL > 160mg/dL tending toward a lower instantaneous risk of death.

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?*

Our analyses suggest a trend toward improved survival with higher levels of LDL. This goes against what we could expect to see as LDL, or “bad” cholesterol, is usually associated with increasing risk for cardiovascular disease and most medical and lifestyle interventions are aimed at reducing LDL. Again, I would have expected a positive association between LDL and 5 year all-cause mortality. In choosing an analytic method a prior, I would have assumed that the data could be heteroskedastic as the mean LDL could vary across many different variables such as age and presence of other comorbidities. With this in mind, simple linear regression would not be able to definitively detect differences between groups vs. rejecting the null hypothesis. Heteroskedastisticy could produce biased estimates with logistic regression. For these reasons, a geometric mean t-test or the hazard ratio could be useful measures of association.