**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #1**

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.

Methods: First I changed the observation time from days into years by dividing the variable by 365.23. Then I looked at the summary variables (minimum, maximum, mean) of observation time among those who died during the study compared to those who did not die during the study period and their observation time includes total time on study.

Inference: The minimum time on study for participants who did not die during the study period is 5.002 years. Therefore it is valid to dichotomize time to death according to death within 5 years of study enrolment or death after 5 years because we know that people who were censored because did not die within the first 5 years.

|  |  |  |  |
| --- | --- | --- | --- |
|  | n | Minimum (years) | Maximum (years) |
| Participant Died during the Study | 133 | 0.186 | 5.54 |
| Participants did not die during the study (administratively right censored) | 602 | 5.002 | 5.911 |

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.

Methods: I created the below table by dichotomizing serum LDL based on high levels (LDL > 160 mg/dL) and normal levels (LDL<160 mg/dL). I then looked at several descriptive statistics of over covariates stratified by serum LDL.

Inference: See table and explanations below.

Table 2 shows that there does seem to be a difference in time of death between high and normal serum LDL levels with 17% of participants dying after 5 years in the normal levels but only 13.7% dying after 5 years in the higher LDL levels. When comparing other variables participants with normal LDL levels weigh less, smoke more, are more likely to have a congestive heart failure diagnosis and less likely to have a cerebrovascular event than those with high serum LDL levels.

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.

Methods: First I dichotomized death by within five years and after five years which was the outcome for subsequent questions dealing with dichotomized five year mortality. A two sample t-test with unequal variance was used to compare mean serum LDL across the binary variable of all-cause mortality before and after 5 years. From this we can find the mean difference, p-value and 95% confidence intervals. The null hypothesis is that the mean serum LDL amongst those dying less than 5 years compared to those dying more than 5 years is the same while the two-sided alternative hypothesis is that the means are different.

Inference:The mean serum LDL among those dying within 5 years of study enrollment is 118.7 mg/dL and the mean serum LDL among those dying after 5 years of study enrollment is 127.2 mg/dL. The difference in mean LDL amongst the 2 survival groups is 8.5 mg/dL with those dying after 5 years of study enrollment having a higher serum LDL. Based on a two-sided p-value of 0.0186, these results are statistically significant and the 95% confidence interval indicates the observed difference falls among a true difference in the population anywhere between 1.44 mg/dL and 15.56 mg/dL. We will reject the null hypothesis and conclude that death after 5 years of study enrollment is associated with a higher mean serum LDL.

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.

Methods: Serum LDL was log-transformed and then a two sample t-test with unequal variance was used to compare geometric mean serum LDL across the binary variable of all-cause mortality before and after 5 years. From this we can find the geometric mean difference, p-value and 95% confidence intervals. The geometric mean difference and 95%CI were exponentiated to compare results. The null hypothesis is that the geometric mean serum LDL amongst those dying less than 5 years compared to those dying more than 5 years is the same while the two-sided alternative hypothesis is that the geometric means are different.

Inference:The geometric mean serum LDL among those dying within 5 years of study enrollment is 112.01 mg/dL and the geometric mean serum LDL among those dying after 5 years of study enrollment is 122.83 mg/dL. The geometric mean serum LDL is 9.65% higher in those who die after 5 years compared to those who die within 5 years of study enrollment. Based on a two-sided p-value of 0.0128, these results are statistically significant and the 95% confidence interval indicates the observed difference falls among a true difference in the population anywhere between 2.0% and 17.9% higher in those who died after 5 years of enrollment. We will reject the null hypothesis and conclude that death after 5 years of study enrollment is associated with a higher geometric mean serum LDL.

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).

Methods: First I dichotomized serum LDL according to less than 160 mg/dL(normal) and LDL > 160 mg/dL (high). This cutoff is used in subsequent questions categorizing serum LDL. In order to compare probability or risk of death within 5 years across high v normal LDL, we can use a difference of proportions to obtain risk difference estimates and 95% confidence intervals and a chi-squared test can be used to test independence. Here the null hypothesis that death within 5 years and high serum LDL are independent and we test the alternative hypothesis that they are not independent.

Inference: Among those with normal serum LDL (<160 mg/dL), the probability of death within the first 5 years after study enrollment was 16.99%. Among those with high serum LDL (≥160 mg/dL), the probability of death within the first 5 years after study enrollment is 13.68%. The difference in the probability of death is 3.3% with those who have normal serum LDL at a higher risk for death than those with high serum LDL. The observed difference would not be unusual if the true difference in the population fell between 10.2% lower risk in high LDL participants and 3.6% higher risk in high LDL participants. Based on a p-value of 0.375 we fail to reject the null hypothesis that death within 5 years and high serum LDL are independent and there is no difference in the risk or probability of death based on serum LDL status.

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).

Methods: An odds ratio can be used to compare the odds of death within 5 years across groups of LDL defined as high and normal. This will give us an estimate as well as a 95% confidence interval and corresponding p-value to test the null hypothesis if the odds of death before or after 5 years is the same. The confidence interval will show whether what the typical values are in the true population and how they differ from an odds ratio of 1.

Inference: When comparing the odds of death within 5 years across high v. normal serum LDL, the odds ratio is 0.774 or the odds of death within 5 years among high serum LDL is 22.6% lower than in those with normal serum LDL. The 95% confidence interval indicates that the observed odds ratio is typical if the true odds ratio in the population is between 0.44 and 1.37 or the odds of death within 5 years among high serum LDL is anywhere between 56% lower or 37% higher compared to normal serum LDL. Based on a p-value of 0.3753 we fail to reject the null hypothesis that death within 5 years and high serum LDL are independent.

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).

Methods: Using survival analysis methods using time on study and the death indicator variable, the logrank test can be used to assess the null hypothesis that the survival function in normal v. high LDL are the same and the alternative hypothesis that these two survival functions are not equal (or are proportional). We can also use proportional hazards regression methods to obtain 95% confidence intervals for the difference of these survival curves. We can also use robust measurements on the regression model. Finally we can show a graph that compares the survival experiences of both high and normal serum LDL participants.

Inference: Based on a two-sided logrank p-value of 0.2664 we fail to reject the null hypothesis that the survival experience in participants with high serum LDL compared to normal serum LDL is the same. Similarly, using regression methods we obtain a hazard ratio of 0.75 or the instantaneous risk of death is 25% lower in those with high serum LDL. The 95% confidence interval suggests this ratio would not be unusual if the true hazard ratio was anywhere between 0.45 and 1.25 or if the instantaneous risk of death was anywhere between 55% lower in high serum LDL and 25% higher high serum LDL. The regression method provides a p-value of 0.268 and we would again fail to reject the null hypothesis that the survival experience in participants with high serum LDL compared to normal serum LDL is the same. The robust model gives very similar estimate for the 95% confidence interval and p-value. Finally the graph below also shows that the curves are similar which support failing to reject the null hypothesis.

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?

We have already established that dichotomizing survival into less than 5 years and greater than 5 years is valid; I would only use measures where we were able to use serum LDL as a continuous measure so we don’t lose important information. This leaves using two-sample t-tests to compare arithmetic mean or geometric mean. Upon comparing the t-statistics the geometric mean t-test shows a strong association (t=2.52) compared to the arithmetic mean t-test (t=2.38). Therefore, the geometric mean shows a stronger association and would be the best method in order to detect as association between mortality and serum LDL.