Biostat 515/518 Homework 1

2014 Winter

**TOTAL GRADE: 37/75**

**Question 1**:

Dichotomizing the time to death to within 5 years of study enrollment or more than five years from study enrollment is valid in this study. This is due to the fact that all deaths that were not observed (aka censored) occurred at least 5 years after that patient was enrolled (the date of their MRI). Patients who had not died on September 16, 1997 were censored, but the patient with the least number of days between his/her MRI and this date was enrolled for 1827 days, which is a couple days past five years. Therefore, dichotomizing in this way removes the issue of censored data since we know with certainty in which of the two groups every patient should belong. 5/5

**Question 2**:

Tables 1 and 2 below summarize descriptive statstics for our sample in this study. More participants had LDL levels below 160mg/dL than above that mark (618 vs 107). The proportion who died within 5 years wsa greater for those with high LDL in all sub-categires except for those who had been diagnosed with angina, myocardial infarction, or congestive heart failure.

Within both high and low LDL groups, those who died within five years were younger and smoked fewer pack-years, on average.

Table 1: Categorical Variables

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **High LDL (>160 mg/dL)** | **Low LDL (<160 mg/dL)** |
|  |   | **N** | **% Died within 5 Years** | **N** | **% Died within 5 Years** |
| Overall |  |  107 | 86.92% | 618 | 83.00% |
| Sex | **Males** | 45 | 80.00% | 315 | 78.41% |
|  | **Females** | 62 | 91.93% | 303 | 87.79% |
| CHD | **No Diagnosis** | 86 | 93.02% | 488 | 86.07% |
|  | **Diagnosis of Anigna** | 8 | 62.50% | 54 | 74.07% |
|  | **Diagnosis of Myocardial Infarction** | 13 | 61.54% | 76 | 69.73% |
| CHF | **No Diagnosis** | 104 | 88.46% | 581 | 84.51% |
|  | **Diagnosis of Congestive Heart Failure** | 3 | 33.33% | 37 | 59.46% |
| Stroke | **No Diagnosis** | 87 | 89.66% | 541 | 85.95% |
|  | **Diagnosis of a Transient Ischemic Attack** | 6 | 83.33% | 18 | 66.66% |
|  |  **Diagnosis of a Stroke** | 14 | 71.42% | 54 | 61.02% |

Table 2: Continuous Variables

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **High LDL (>160 mg/dL)** | **Low LDL (<160 mg/dL)** |
|  |  | Died within 5 Years | Survived 5 Years | Died within 5 Years | Survived 5 Years |
| Age (yrs) | Mean(sd) | 74.46(5.46) | 77.64(7.14) | 74.12(5.17) | 76.44(6.05) |
|  | Min-Max | 65-94 | 69-89 | 65-99 | 67-91 |
| Smoking History (Pack-years) | Mean(sd) | 16.66(23.58) | 27.55(28.50) | 18.22(24.93) | 28.03(37.20) |
|  | Min-Max | 0-102 | 0-78 | 0-180 | 0-240 |

While mean and standard deviation are presented for the continuous variables, the following measures are missing: the sample size and the number of missing observations. The formatting of the table is clear, and units of measurements are listed.

General table layout and design: 4/4. Tables are well designed and well labeled.

Choice of descriptive statistic: 1.5/3 Points deducted for not including sample size and missing observations on continuous variables. In addition, in your table of categorical variables it is unclear how the counts are performed. IE, is N the number of total people in that LDL group with a diagnosis of angina etc. or is it related to the percent of people with angina in the LDL group who died within 5 years. This is unclear.

Discussion of the finding: 1.5/3. Your discussion of the descriptive statistics lacks a statement about any missing data, or any effects of missing data. It also lacks a discussion of how sex relates to the different LDL groups, and statements about whether there are consistent trends across the data.

Total: 7/10

Note: Discussion of the methods used to perform this analysis is missing.

**Question 3**:

A two-sample t-test was performed to compare mean LDL levels of those who did and did not survive past 5 years of their MRI date. Did this t test assume equal variances, or did it allow for unequal variances? What was the assumption on the creation of your 95% CI?. There is evidence that the mean LDL level is different between those who did and those who did not survive past five years after their MRI (p-value of 0.0186). Those who died had an average LDL 8.501mg/dL higher than those who survived. With 95% confidence the true difference in arithmetic mean is between 1.441 and 15.697 mg/dL.

Methods: 3.5/5. Analysis (t-test) is correct. However a better discussion of the assumptions of the t test used is required, as is a description of whether the test is two-sided or one-sided.

Reporting of Association: 4/5. Although you reported the difference in the average LDL’s, you should also state what the mean LDL levels are for each group (IE what is the summary measure for each distribution). Then you can state how you are comparing the distributions (difference in those means, or the ratio of those means etc.)

Total: 7.5/10

**Question 4**:

A two-sample t-test was performed to compare the geometric mean LDL levels of those who did and did not survive past 5 years of their MRI date. This was accomplished by taking the log of the LDL levels, performing the test, and exponentiating both point the estimate and the confidence interval bounds. There is evidence that the geometric mean LDL level is different between those who did and those who did not survive past five years after their MRI (p-value of 0.0128). Those who died had n geometric average 1.0965mg/dL higher than those who survived. With 95% confidence the true difference in geometric mean is between 1.020 and 1.1787 mg/dL.

Methods: 4/5. As before, it’s important to mention the assumptions of your t-test, and that those assumptions hold for the creation of the 95% CI (are you assuming equal variances, or are you allowing for unequal variances?)

Reporting of Association: As above in question 3, you need to mention not just the difference in the geometric means, but what the absolute values are for each sample. Your p-value should also state more explicitly whether this is a one sided or a two sided test. Because you didn’t state the geometric means for each group, it is hard to tell whether your difference in the geometric means has been back-transformed or not. 3.5/5

Total: 7.5/10

**Question 5**:

A two-sample t-test was performed to compare the probability of death within five years between those with high and low LDL levels, where “high” is defined to be >160mg/dL. There is no evidence that the probability of death is different for people with high LDL levels (>160mg/dL) versus low LDL (p-value of 0.2806). Those with high LDL levels had a probability of death within five years 0.039mg/dL higher than those with low levels of LDL. With 95% confidence the true difference in probability is between -0.0322 and 0.1103 mg/dL.

Methods: Here, the t test is not the best test to use. Since you have a binary POI (high or low LDL) and a binary outcome (live < or > 5 years), a better choice would be a chi-square test, to test whether having high or low LDL is independent of the survival outcome. You also need to state what method was used for computing your 95% CI (Wald is standard) 0/5.

Reporting of Association: You have reported your summary measure as being a difference in the concentration of LDL between two groups, yet you call this a representation of the probability of death. It would be helpful for you to think about what is actually needing to be measured. Here, we are not interested in the differences in the groups’ LDL concentrations, but rather in the survival probabilities that each group has. 0/5

Total:0/10

**Question 6**:

The probability of dying within 5 years for participants with high levels of LDL is estimated to be 0.8692, while the probability for the low LDL group is estimated to be 0.8301. This leads to odds of 6.643 and 4.886, respectively, and an odds ratio of 1.359. In other words, participant with high LDL levels (at least 160mg/dL) were 1.359 times as likely to die within five years as those with high LDL levels (at least 160mg/dL) were 1.359 times as likely to die within five years as those with low levels of LDL. An odds ratio test suggests that the true odds ratio is between 0.7349 and 1, with a 95% confidence level.

Methods: You should talk about how you derived the probability of dying and how you used those numbers to compute the odds of dying (ie, it appears that you’ve calculated the odds by doing odds/1-odds, but you could also have run a statistical test. How you get the odds and the odds ratio is important. Also, how are you computing your confidence intervals (ie Wald or exact methods?). While you have reported numbers, you have not stated any methodology regarding how you derived those numbers 0/5.

Reporting of Association: Your reported odds ratio doesn’t fall within your 95% confidence interval. This should wave a red flag since it demonstrates that your odds ratio and your confidence intervals are not related. You have not stated a p value either on your estimate of the odds ratio, and therefore we don’t know how reliable that ratio is. Points have been awarded since you did mention what your calculated odds were for both groups, and what the ratio of those odds is. 2/5

Total: 2/10.

**Question 7**:

A Cox regression comparison of the instantaneous risk of death (hazard) between the high and low LDL groups yields a hazard ratio of 0.7179. There is no evidence that these instantaneous risks are different (p-value of 0.227). A 95% confidence interval suggests the true hazard ratio is between 0.4193 and 1.229.

Methods: You still need to scientifically define who the high and low ldl groups are (ie, what is the threshold for dichotomization). You also need to define how your 95% CI was calculated. Also, the survival distribution must be estimated using Kaplan-Meier methods first before you can compare those numbers to determine the hazard ratio between the two groups. 0/5

Reporting of Association: While your p-value is accurate, your interpretation of it is wrong. The low p-value does not mean that there is no evidence whatsoever of an association; what it means is that your observed values would not be unlikely even if there was no association, and that can be affected by a variety of factors such as sample size, test precision etc. that do not pertain to evidence of an association. Your explanation of the 95% confidence interval is also lacking nuance…it’s not that 95% of the time the true mean will be within those parameters, rather the 95% CI represents a range of hazard values *given your data* that the true mean could assume. In addition to the hazard ratio, you need to talk about the instantaneous probability of death for each group, which you have not mentioned. Points have been awarded because you did use cox regression for proportional hazards. 2/5

Total: 2/10

**Question 8**:

To answer the question about an association between mortality and serum LDL, a priori I would have chosen to look at the hazard ratio, as it is the most efficient in the sense that we need not lose information by categorizing participants into those who died and those who did not die within five years. We are able the more precise measurements of time to death (or censoring).

2 points for mentioning efficiency of not dichotomizing. 4 points for providing a test that is aligned with the aim of efficiency.

6/10

TOTAL GRADE ON ASSIGNMENT: 37/75