**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #1**

January 6, 2014

30/40

All questions relate to associations between death from any cause and serum low density lipoprotein (LDL) levels in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. The data is in free-field format, and can be read into Stata using the following code in a .do file.

infile ptid mridate age male race weight height packyrs yrsquit alcoh ///

physact chf chd stroke diabetes genhlth ldl alb crt plt sbp aai ///

fev dsst atrophy whgrd numinf volinf obstime death ///

using http://www.emersonstatistics.com/datasets/mri.txt

Recommendations for risk of cardiovascular disease according to serum LDL (low density lipoprotein) levels are as follows (taken from the Mayo Clinic website):

|  |  |
| --- | --- |
| Below 70 mg/dL | Ideal for people at very high risk of heart disease |
| Below 100 mg/dL | Ideal for people at risk of heart disease |
| 100-129 mg/dL | Near ideal |
| 130-159 mg/dL | Borderline high |
| 160-189 mg/dL | High |
| 190 mg/dL and above | Very high |

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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Obs** | **Mean** | **Std. Dev.** | **Min** | **Max** |
| Observation time (years) | 602 | 5.327021 | 0.296818 | 5.002053 | 5.91102 |

Table 1: Observation time (in years) for patients with right-censored observations

The first censoring event occurred after 5 years of study enrollment (the minimum observation time until censoring is 5.002 years), so dichotomizing the time to death according to death within 5 years of study enrollment is a valid approach.

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.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | High LDL (>160mg/dL) | | | Low LDL (<160mg/dL) | | | Total | | |
|  | N (nmissing) | Mean (sd) | Min-max | N (nmissing) | Mean (sd) | Min-max | N (nmissing) | Mean (sd) | Min-max |
| Age (years) | 622 (0) | 74.7 (5.6) | 65-94 | 113 (0) | 74.5 (5.4) | 65-99 | 735 (0) | 74.6 (5.5) | 65-99 |
| Weight | 622 (0) | 163.5 (30.8) | 74-257 | 113 (0) | 159 (31) | 86-264 | 735 (0) | 160 (30.7) | 74-264 |
| Pack years | 622 (0) | 18.6 (24.5) | 0-102 | 112 (1) | 19.8 (27.6) | 0-240 | 734 (1) | 19.6 (27.1) | 0-240 |

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | High LDL | Low LDL | Total |
| Male | 51/117 | 315/618 | 366/735 |
| Congestive heart failure | 0: 581/618  1: 37/618 | 0: 113/117  1: 4/117 | 0: 694/735  1: 41/735 |
| Congenital heart disease | 0: 488/618  1: 54/618  2: 76/618 | 0: 92/117  1: 10/117  2: 15/117 | 0: 580/735  1: 64/735  2: 91/735 |
| Stroke | 0: 541/618  1: 18/618  2: 59/618 | 0: 95/117  1: 6/117  2: 16/117 | 0: 636/735  1: 24/735  2: 75/735 |

4/4 for general table layout

1/3 for the choice of descriptive statistics

0/3 for discussion of finding

Did not mention about missing data (-1)

Wrong choice of descriptive statistics e.g. binary variable only takes mean(-1)

Total: 5/10

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.

Method: A two sample t-test of serum LDL between groups defined by vital status at 5 years was performed to evaluate an association between serum LDL and 5 year all-cause mortality. Since the problem does not presume equal variances between the two groups, the t-test was performed allowing for unequal variances.

The mean serum LDL levels for the group of subjects who died within 5 years and are alive at 5 years are 118.7 mg/dL and 127.2 mg/dL respectively, with a difference of 8.5 mg/dL across groups defined by vital status. The result is significantly different from 0 (P=0.0186), with a 95% confidence interval suggesting that the observed difference is what might be typically observed if the true difference between 5 year survivors was anywhere between 1.44 mg/dL and 15.56 mg/dL, with the survivors averaging higher cholesterol. We thus reject the null hypothesis of no association between serum LDL and 5 year all-cause mortality in favor of a trend toward higher mean cholesterol for the subjects surviving for a longer period of time.

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.

Method: To compare the geometric mean LDL values across groups defined by vital status at 5 years, a two-sample t test on the log transformed serum LDL levels was performed, again allowing for unequal variances between the groups. Point estimates and confidence intervals were back transformed to obtain estimates in the original measurement units.

The geometric mean cholesterol is estimated to be **112** mg/dL among subjects who die within 5 years of study enrollment and **123** mg/dL among subjects who survive at least 5 years. A comparison of the geometric means estimates that the geometric mean cholesterol is **9.7%** higher among subjects who survive at least 5 years relative to those who die within 5 years. The observed difference in geometric mean LDL values is beyond that that might be expected to occur by chance in the absence of a true effect of serum LDL on 5 year all-cause mortality (**P=0.0128**), with the 95% confidence interval suggesting that the observed difference is what might be typically observed if the true ratio of geometric means was anywhere between **2%** and **17.9%**, with the survivors averaging higher cholesterol. We thus reject the null hypothesis of no association between serum LDL and 5 year all-cause mortality in favor of a trend toward higher geometric mean cholesterol for the subjects surviving for a longer period of time.

5/5 for performing an appropriate analysis

5/5 for reporting the association appropriately

Total: 10/10

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).

|  |  |  |  |
| --- | --- | --- | --- |
|  | High LDL  (>160 mg/dl) | Low LDL  (<160 mg/dL) | Total |
| Death within 5 years | 14 | 105 | 119 |
| Surviving at 5 years | 93 | 513 | 606 |
| Total | 107 | 618 | 725 |

Table 2: 2x2 table grouped by LDL status (high LDL being >160 mg/dL) and vital status at 5 years.

Method: A Chi-square test on the two groups defined by LDL status (defined as >160 mg/dL) and 5 year all-cause mortality was performed to evaluate the null hypothesis that the risk difference of 5 year all-cause death between groups defined by LDL status equals zero.

In the group of patients with high LDL (serum LDL >160 mg/dL), 14 of 107 subjects (13%) died within 5 years of study enrollment and 105 of 618 subjects (17%) with low LDL died within 5 years. Based on the Chi squared test, the observed absolute difference of -3.9% in probability of death within 5 years is within the range of absolute differences that might be expected to occur by chance in the absence of an effect of high serum LDL on 5 year all-cause mortality (P=0.3139), with the 95% confidence interval for the difference in 5 year death rates is an absolute 11% lower to 3% higher 5 year death rate in the high LDL group.

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).

Method: A Chi-square test on the two groups defined by LDL status (defined as >160 mg/dL) and 5 year all-cause mortality was performed to evaluate the null hypothesis that the odds ratio of 5 year all-cause death between groups defined by LDL status equals 1.

The odds of dying within 5 years is estimated to be 26.4% lower (odds ratio 0.735) in the group having high LDL relative to the group with low LDL. This observed difference is not statistically different from an odds ratio of 1 (P=0.3139), with a 95% confidence interval suggesting that the observed odds ratio is what might be typically observed if the true odds of dying within 5 years was anywhere between 63% lower and 36% higher (odds ratios 0.3726 and 1.3606, respectively) in the high LDL group relative to the low LDL group. Therefore we do not reject the null hypothesis of no association between serum LDL and survival time, even though there is a slight trend towards higher odds of survival among subjects with higher serum LDL levels.



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).

Method: A log-rank test comparing the survival curves of the two groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL) to test the null hypothesis that the survival curve for patients with high serum LDL is the same as the patients with low serum LDL.

|  |  |  |
| --- | --- | --- |
|  | Events | Events |
| LDL status | observed | expected |
| High | 116 | 111.01 |
| Low | 15 | 19.99 |
| Total | 131 | 131 |

Table 3: Number of observed and expected deaths for groups separated by LDL status (“high” = LDL > 160 mg/dL)

The log-rank test results indicate there is not a significant difference in the survival curves between groups defined by whether the subjects have high serum LDL (P=0.2249). Thus, we do not reject the null hypothesis that the instantaneous risk of death in the high serum LDL group is equal to that of the low serum LDL group. The figure below shows the significant overlap between the 95% confidence intervals of the two groups over a majority of the study duration.



Figure 1: Kaplan-Meier survival estimates and 95% confidence intervals for groups defined by high serum LDL (“high”= LDL>160mg/dL)

5/5 for performing an appropriate analysis

4/5 for reporting the association appropriately

Did not report whether the p-valu is two-sided or one-sided(-1)

Add KM(+1)

Total: 10/10

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?

I would have preferred the analysis comparing the instantaneous risk of death across groups defined by whether the subjects have high serum LDL. Given that the observation times are right-censored, the survival-based method is most able to leverage the information from the censored observations.

. Performed analysis that are valid (2)

Appropriate access(3)

Total: 5/10