**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #1**

January 6, 2014

**29.5/40**

**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.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***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.***

*Keys to past homeworks from quarters that I taught Biost 517 (e.g. HW #8 from 2012) or Biost 518 (e.g., HW #3 from 2008) or Biost 536 (e.g. HW #3 from 2013) might be consulted for the presentation of inferential results. Note that the requirement to provide a paragraph describing your statistical methods is new this year, and thus past keys do not give explicit examples of a separate paragraph. However, many past keys provide this information as an introductory sentence.*

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

Note that the first line of the text file contains the variable names, and will thus be converted to missing values. Similarly, there is some missing data recorded as ‘NA’, and those, too, will be converted to missing values. If you do not want to see all the warning messages, you can use the “quietly” prefix. You may want to go ahead and drop the first case using “drop in 1”, because it is just missing values.

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.
* When separate observed time to two groups, longer than 5 years or not, we can see death status of everyone in the less than 5 years group is death. That means we can observe every person’s time in this group. The frequency table is the following: all people in within 5 years group are died. Therefore, we can use whether those people survive longer than 5 years as an indicator to show those people’s status at 5 years.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Within 5 years | After 5 years | Total |
| Non-Death | 0 | 602(98.05%) | 602 |
| Death | 121(100%) | 12(1.95%) | 133 |
| Total | 121 | 614 | 735 |

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.
* Smoking (45 versus 33 pack years), male (64 % versus 47%), having diagnosis of congestive heart failure prior to MRI (14 % versus 4%) myocardial infarction (24 % versus 10%), and stroke (23% versus 8%) is more likely to die within 5 years. Female (56 % versus 49%), stroke (13% versus 10%), and weight (163 versus 159 pounds) is more likely to have higher LDL.
* Congestive heart failure, coronary heart disease, and cerebrovascular event might be the causal pathway from LDL to be died within 5 years, so we do not consider these variables are confounder. Gender and smoking might be a confounder even though the dataset did not show too much that smoking is associated with LDL.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 　 | Death within 5 yearsN (%) | Death after 5 yearsN (%) | LDL >160 N(%) | LDL $\leq $ 160N(%) |
| Gender |  |  |  |  |
| Male  | 78 (64.46) | 288 (46.91) | 50 (44.25) | 316 (50.80) |
| Female | 43 (35.54) | 326 (53.09) | 63 (55.75) | 306 (49.20) |
| Congestive heart failure prior to MRI  | 　 | 　 | 　 | 　 |
| Yes | 17 (14.05) | 24 (3.91) | 4 (3.54) | 37 (5.95) |
| No | 104 (85.95) | 590 (96.09) | 109 (96.46) | 585 (94.05) |
| Coronary heart disease prior to MRI  |  |  |  |  |
| Myocardial infarction | 29 (23.97) | 62 (10.10) | 14 (12.39) | 77 (12.38) |
| Angina | 17 (14.05) | 47 (7.65) | 10 (8.85) | 54 (8.68) |
| No | 75 (61.98) | 505 (82.25) | 89 (78.76) | 491(78.94) |
| Cerebrovascular event prior to MRI | 　 | 　 | 　 | 　 |
| Stroke | 28 (23.14) | 47 (7.65) | 15(13.27) | 60 (9.65) |
| Transient ischemic attack | 7 (5.79) | 17 (2.77) | 6 (5.31) | 18 (2.89) |
| no | 86 (71.07) | 550 (89.58) | 92 (81.42) | 544 (87.46) |
| Smoke |  |  |  |  |
| Yes | 76 (62.81) | 338 (55.05) | 64 (56.64) | 350 (56.27) |
| No | 45 (37.19) | 276 (44.95) | 49 (43.36) | 272 (43.73) |
|  |  |  |  |  |
|  | Mean(SD) | Mean(SD) |  |  |
| age | 76.48 (6.17) | 74.19 (5.22)  |  74.69 (5.63)  |  74.54 (5.42)  |
| smoking (pack year) | 44.87 (36.38)  | 32.61 (25.10) |  33.07 (24.27)  |  35.15 (28.46)  |
| weight (pounds) | 159.12 (32.79) | 160.11 (30.35)  |  163.47 (30.81)  |  159.31 (30.71)  |

4/4 for general table layout

2/3 for the choice of descriptive statistics

3/3 for discussion of finding

Did not mention about missing data (-1)

Total: 9/10

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.

For problem 2-6, we use whether the observed time longer than 5 or not as an indicator to define the death status at 5 years. If the observed time is less than 5 years, then death status =1; if the observed time is longer than 5 years, then death status=0.

Our potential of interest (POI), LDL, is a continuous variable. Thus we can analyze data by using two-sided, unequal variance two-sample t-test to compare the mean of LDL between two groups. The hypothesis is the following:

Let $μ\_{A}=mean LDL in the group that survival time less than 5 year$

$$μ\_{B}=mean LDL in the group that survival time longer than 5 year$$

H0: $μ\_{A}=μ\_{B}$

H1: $μ\_{A}\ne μ\_{B}$

From the following result, we see the mean difference of LDL is – 8.5 mg/dL. With 95% confidence, it would not be surprised if the true mean difference of LDL between (-15.1, -1.9) mg/dL. We reject the null hypothesis because of the p-value = 0.019, that is, mean LDL in two groups is different.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N | Mean | SD | 95% CI |
| Within 5 years | 119 | 118.7 | 36.2 | (112.1, 125.3) |
| After 5 years | 606 | 127.2 | 32.9 | (124.6, 129.8) |
| Mean diff. |  | -8.5 | 3.4 | (-15.1, -1.9) |

t = -2.38, p-value = 0.019

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.

As mentioned above, LDL is a continuous variable, and we can use log transformation for comparing geometric mean between two groups. First, we create a new variable by log transformation for LDL, and then repeat the same procedure for two-sided, unequal variable two-sample t-test. Finally, we transform our mean, and 95% CI back by exponentiation. The mean different of LDL is 1.1 mg/dL, and 95%CI means that it is not unusual if the mean different between 1.0 and 1.2 mg/dL. With p-value = 0.013, we reject the null hypothesis. The mean LDL in two groups is different.

|  |  |  |  |
| --- | --- | --- | --- |
|  | N | Mean | 95% CI |
| Within 5 years | 119 | 122.8 | (120.2, 125.5) |
| After 5 years | 606 | 112.0 | (104.5, 120.0) |
| Mean diff. |  | 1.1 | (1.0, 1.2) |

 t = 2.52, p-value= 0.013

5/5 for performing an appropriate analysis

2.5/5 for reporting the association appropriately

Wrong report the point estimate(ratio of geometric mean) (-1) mean different of LDL => the ratio of GM is 1.1

Did not report which of geometric mean of LDL between two groups is higher (-0.5)

Wrong interpretation of CI (-1)

Total: 7.5/10

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).

After separating LDL to two groups, we have two categorical variables, and expected value of each cell in the following frequency table is larger than 5, so we can use chi-square test.

H0: LDL and whether death at 5 year is independent

H1: LDL and whether death at 5 year is not independent

The p-value is 0.32, not reject the null hypothesis. We don’t have enough evidence to prove LDL and death at 5 year is not independent.

|  |  |  |  |
| --- | --- | --- | --- |
|  | LDL $\geq $ 160 | LDL < 160 | Total |
| Within 5 years | 15 | 106 | 121 |
| After 5 years | 98 | 516 | 614 |
| Total | 113 | 622 | 735 |

 $χ^{2}= 0.987$ , p-value =0.32

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).

Exposure: LDL$ \geq $ 160 mg/dL

Case: death within 5 years = 1

The odds of case among high LDL group is 0.133, and the odds of case among low LDL group is 0.17. The risk difference is – 0.037, 95% CI is (-0.107, 0.032). With 95% confidence, it is not surprised if the true risk difference between -0.107 and 0.032. Calculate the odds ratio = 0.745, 95% CI: (0.419, 1.325). That is not unusual if the true odds ratio is between 0.419 and 1.325.

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).

Because of censored data, we should use log-rank test to compare the hazard rate in the high and low LDL group.

H0:$ h\_{1}(t)=h\_{2}(t) for all t$

H1: $h\_{1}(t)\ne h\_{2}(t)$

We get the chi-square statistics with 1 degree of freedom is 1.44, p-value =0.2301, fail to reject the null hypothesis. That is, we do not have enough evidence to prove the hazard rate between high and low groups is different.

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)

Total: 9/10

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 prefer using survival analysis to discuss if mortality is associated with serum LDL because the data is right censor. It is also better to divide serum LDL to high and low level because it would be easier to see if any different survival time between these two groups. Thus, we can compare whether the survival time is different between high and low serum LDL. The null hypothesis assumed no different of survival time between two groups.

Performed analysis that are valid (2)

Log-rank test+censored data (2)

Total 4/10

**Discussion Sections: January 6 – 10, 2014**

We will review material from Biost 517 / 514 as it relates to the scientific question posed by this homework. Come to discussion section prepared to discuss (and ask questions) about this assignment.