**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #1**

January 6, 2014

**Total point:49**

**Written problems:** To be submitted as a MS-Word compatible email attachment to [semerson@uw.edu](mailto: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.

**Method: Analysis of the minimum and the maximum observation time between those who died within the study and those who did not, can provide justification for dichotomization of observation time into less than or greater than five years to describe time to death events while seemingly disregarding censoring.**



**Table 1.** Description of observation time in years. Category Death=0 indicates death was not observed during the study. Death=1 indicates death was observed during the study.

**Inference: The minimum observation time recorded for subjects whose death was not observed during the study was slightly larger than five years (5.002 years). Thus, when dichotomizing observation time into less than of greater than five years, all subjects in the category of ≤5 years died within 5 years of study time, and most of those who passed the five year mark survived for the entire study.**

**5/5**

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.

**Method: Descriptive statistical analysis was performed on variables, comparing across high LDL and low LDL groups.**

**Inference: Summary measures presented for age, weight, heart failure, heart disease, stroke and smoking history in Table 2, appear to be very similar across LDL groups. Proportion of subjects with and without a death event was similar across LDL groups.**



**Table 2.**

4/4 for general table layout and labeling of columns, rows, and descriptive statistics

3/3 for choice of descriptive statistics

0/3 for discussion of the finding

Total = 7

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 two sided t test was performed comparing mean LDL values across groups defined by vital status at or before 5 years in the study (all subjects were observed to die), and after 5 years (subjects survived the length of the study). Values were computed using non-log transformed data. CI were calculated assuming unequal variance and an alpha level of 0.05.**

**The null hypothesis: Mean LDL values are similar/same between those who died during the study and those who survived during the study.**

**Alternative hypothesis: Mean LDL values are different between those who died during the study and those who survived during the study**



**Table 3.**

**Inference: The results showing that the average LDL value for those who survived during the study is larger than those who died by 8.5 mg/dL would not be surprising if the true mean LDL is greater in those who survived than those who died by 1.44 to 15.56 mg/d. Given the p value of 0.0186 the null hypothesis can be rejected.**

5/5 for performing an appropriate analysis and describing the method appropriately

4/5 for reporting the association appropriately

did not mention how those groups are being compared

Total = 9

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: A two sample two sided t test was performed comparing mean LDL values across groups defined by vital status at or before 5 years in the study (all subjects were observed to die), and after 5 years (subjects survived the length of the study). Values were computed using natural log transformed data. CI were calculated assuming unequal variance and an alpha level of 0.05.**

**The null hypothesis: Mean LDL values are similar/same between those who died during the study and those who survived during the study.**

**Alternative hypothesis: Mean LDL values are different between those who died during the study and those who survived during the study**



**Table 4.** mean values and standard deviation were calculated by exponentiating the log value that was used to compute the t test.

**Inference: Given the p value of 0.0128, the null hypothesis can be rejected.**

5/5 for performing an appropriate analysis and describing the method appropriately

2/5 for reporting the association appropriately

Hypothesis is wrong

Did not mention

1. How those groups are being compared

2. No explanation about point estimate, CIs.

Total = 7

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: The probability of death was compared across high and low LDL groups. A chi squared test was performed with the count of people who died or didn’t die with low and high LDL. The probability was calculated as the risk of death, and difference in risk is provided.**



**Inference: given the chi squared p value of 0.375, we fail to reject the null hypothesis that the probability of death is similar among LDL groups.**

5/5 for performing an appropriate analysis and describing the method appropriately

2/5 for reporting the association appropriately

P(death| HighLDL), p-value are wrong

Did not mention

1. No explanation about CIs.

Total = 7

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: The odds of death were compared across high and low LDL groups. A chi squared test was performed with the count of people who died or didn’t die with low and high LDL. The odds ratio was calculate along with 95% intervals**



**Inference: given the chi squared p value of 0.375, we fail to reject the null hypothesis that the odds of death are similar among LDL groups.**

5/5 for performing an appropriate analysis and describing the method appropriately

1/5 for reporting the association appropriately

Odds(death| HighLDL), p-value and OR are wrong

Didn’t do interpretation

Total = 6

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: A logrank test was computed evaluate the association of mortality and low and high LDL. Also a cox te**

**Results: chi square=1.23; p-value= 0.2664**

**Inference: Given the p-value of 0.2666 we cannot reject the null hypothesis that the survival/mortality probability between High and low LDL groups during the entire study is the same**

5/5 for performing an appropriate analysis and describing the method appropriately

3/5 for reporting the association appropriately

p-value is wrong

Total = 8

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? **The logrank test would have been the preferred test *a priori* to answer the question about an association between mortality and serum LDL since it is the test that more appropriately takes into account time to death event over the entire study time.**

**0/10**

**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.