**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #1**

January 6, 2014

**After this first homework, we should all know now the detail of information expected of us. The author of this report has a solid grasp on testing means. However, the author should consider more practice on other tests so that they include more detail.**

**44/80 = 11/20**

**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.

|  |  |  |
| --- | --- | --- |
|  | Death within 5 years of enrolment | Death after 5 years of enrolment |
| Death=1 (true death)  | 121 (100%) | 12 (1.95) |
| Death=0 (censored) | 0 (0%) | 602 (98.05) |

From the data we can see, the right censoring data starts from when the obstime=60.1 month. The survival data within 5 years of study enrolment are the data without censoring. STOP There are 121 (16.5) people died within 5 years of enrolment. There are 614 (83.5%) people survived after 5 years of enrolment. Among the people survived after 5 years of enrolment, most of them (602, 98.05%) survived after at the end of the study. So it is reasonable to dichotomize the time to death according according to death within 5 years of study enrolment or death after 5 years.

 The reasoning is here, but the minimum was not stated as a method. The answer is thorough, but more than necessary. I will take 1 point for not stating the minimum, and another point for using the name of a variable within code (“obstime”). A report should not be written in coding language.

3/5

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.

|  |  |  |
| --- | --- | --- |
| Characteristic | Death within 5 years of enrolmentMean (SD) Min p50 Max | Death after 5 years of enrolmentMean (SD) Min p50 Max |
| Age |  76.5 (6.2) 67 75 91 |  74.2 (5.2) 65 73 99 |
| Weight | 159.1 (32.8) 96 154 264 | 160.1 (30.3) 74 159 258 |
| Smoking in pack years |  28.0 (36.0) 0 18 240  |  17.9 (24.7) 0 4.4 180 |
| Years since quitting smoking |  10.7 (14.4) 0 0 56 |  9.5 (14.0) 0 0 56 |
| Sex \*  Male Female |  78 (64.5)  43 (35.5) |  288 (46.9)  326 (53.1) |
| Congestive heart failure \* No Yes |  104 (86.0) 17 (14.1)  |  596 (96.1) 24 (3.9) |
| Coronary heart disease \* No Angina Myocardial infarction |  75 (62.0) 17 (14.1) 29 (24.0) | 505 (82.2) 47 (7.7) 62 (10.1) |
| Stroke \* No Transient ischemic attack Stroke |  86 (71.1) 7 (5.8) 28 (23.1) |  550 (89.6) 17 (2.8) 47 (7.7) |

\*: For these variables, Mean (SD) is Counts (Percentage)

Compare to the people survived at 5 years of enrollment, the people died within 5 years of enrollment are tend to smoke more, more male than female, more have prior history of cardiovascular disease.

The table has pertinent information and the chosen descriptive statistics are appropriate. Note that missing data was not reported (-1). At first I believed there could have been a formatting issue when I downloaded this report onto my computer, but after checking on school computer, I still find that the table spacing is a bit open and does not lend itself to quick interpretation (-1). Sample size per subset was not reported (-1). It would have been useful to include a combined group, or a third column as shown in Key (-1). Although a summary of the table is given, it is too brief and does not utilize the values of the table (-1).

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.

Two sample two sided t-test was performed to compare mean LDL values across groups defined by vital status at 5 years. The mean LDL value of people died within 5 years of enrollment was 118.70 (95% CI: 112.13 – 125.26). The mean LDL value of people died after 5 years of enrollment was 127.20 (95% CI: 124.57 – 129.83). The resulting t-test p – value was 0.02. With 95% of confidence, the data contained the evidence that the mean LDL value of people died within 5 years of enrollment was lower than those of the people died after 5 years of enrollment.

There is no mention of how the variances were handled (-1). The p-value is incorrect (likely because of the variance issue). The correct p-value is 0.0186. If you rounded up, do not do that. Otherwise, you arrived at the incorrect p-value (-1). A point estimate per sub-group is given, but a point estimate for the difference is not, nor a confidence interval for that difference (-1). The language of the conclusion needs to be more specific to association. A significant difference was found, but the implication is not clear (-1).

6/10

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.

LDL values were log transformed. Then two sample two sided t-test was performed to compare geometric mean LDL values across groups defined by vital status at 5 years. The geometric mean LDL value of people died within 5 years of enrollment was 4.72 (95% CI: 4.65 – 4.79). The mean LDL value of people died after 5 years of enrollment was 4.81 (95% CI: 4.79 – 4.83). The resulting t-test p – value was 0.01. With 95% of confidence, the data contained the evidence that geometric men LDL value of people died within 5 years of enrollment was lower than those of the people died after 5 years of enrollment.

The variance handling was not specified (-1). A point estimate for the difference was not stated (-1). Estimates were not transformed back for interpretation (-1). Incorrect p-value, naturally because variances handling was not done (-1). Mention of the test estimate as a ratio is not explicitly stated (-1).

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).

Chi-Squared text was performed to compare the the probability of death within 5 years across groups defined by whether the subjects have high serum LDL. Based on p-value=0.38, we can not reject the null hypothesis that there’s no association between 5 years mortality and high serum LDL.

This answer seems to have been written in a rush. Is a difference being used? (-1) Is this test 1-sided or 2-sided? (-1) The conclusion is correct. P-value is incorrect (-1). Survival probabilities within group are not reported (-1). The significance level is not reported either (-1), although it does seem unnecessary given the high p-value.

5/10

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).

|  |  |  |
| --- | --- | --- |
|  | Death after 5 years Death within 5 years | Total |
| LDL < 160 mg/dL |  513 105 |  618 |
| LDL >= 160 mg/dL  |  101 16 |  117 |
| Total |  614 121  |  735 |

The odds of death within 5 years in group LDL<160 mg/dL is 105/513=0.20. The odds of death within 5 years in group LDL>=160 mg/dL is 16/101=0.16. The odds ration is 0.16/0.20=0.8. The odds of death within 5 years in the subjects have high serum LDL is 0.8 time less than those of the subjects who have serum LDL lower than 160mg/dl.

The estimated odds of death within group are slightly off. For less than 159 mg/dL, the correct estimate is 0.205, so close enough. However, for high LDL, the estimate is supposed to be 0.151, while the estimate reported is 0.16 (-1). No confidence interval is presented (-1). No test is presented (-1). No p-value is presented (-1). The odds ratio is incorrect, but only because the first two estimates were incorrect. Otherwise, the ratio was correctly estimated. This analysis is more descriptive than inferential (-1).

5/10

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).

Log-rank test was used to evaluating the association. Based on p-value of 0.27, we can not reject the null hypothesis of equal survival probabilities between serum LDL groups.

The correct conclusion is reached. A confidence interval is not presented (-1). An estimate of the test statistic is not provided (-1). The p-value is incorrect (correct = .227) (-1). Since logrank test is a comparison of survival experience, a graph of Kaplan-Meier survival estimates is appropriate (-1).

6/10

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 prefer using logrank test to answer the question about an association between mortality and serum LDL. Because the data are right censored. Logrank test is suitable for censored observations. Logrank test statistic compares estimates of the hazard functions of the two groups. It can use all the information collected including censored data to evaluate the association. It is the most efficient method for this dataset.

An analysis that is valid is stated and from above, the author does know how to perform the test, although not with thorough detail (+2). The drawbacks from other methods are not stated, such as the problem with dichotomizing a continuous measurement or the simplicity of the mean. In considering the log-rank test, the author suggests summarizing the survival distribution (+2). Otherwise, nothing more that is consistent with the Key is presented.

4/10