**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #4**

January 27, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 3, 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.)*

***Unless explicitly told otherwise in the statement of the problem, 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.***

This homework builds on the analyses performed in homeworks #1, #2, and #3. As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex 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. See homework #1 for additional information.

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled as a continuous variable.
	1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.
* Since we are interested in comparing the hazard of death across groups with different LDL levels, which is modeled as acontinuous variable, we used proportional hazard regression with robust standard error. Because it is a censoring distribution, we use Kaplan-Meier estimate to provide the description about the percentiles…... For using K-M method, we dichotomized LDL into two levels: higher than or equal to 160 mg/dL and lower than 160 mg/dL.
* From the result of proportional hazard regression, we get the hazard ratio is 0.9926 and 95% CI is (0.9871, 0.9982). When changed 1 mg/dL unit difference of LDL, the hazard of death is 0.74% lower. With 95% CI, it is not unusual if true hazard of death for the group with1 mg/dL unit higher LDL is from 0.18 lower to 1.29 lower than the group with lower LDL. The p-value is 0.009 < alpha=0.05, has statistically significant.
* From the K-M curve, we can see each line are overlapping in early follow up. After 5 year, the group with LDL >=160 has higher survival, and the group with LDL <100 has lower survival. From the summary table, the incidence rate is decreasing when LDL is increasing. The lowest survival at year 1 is the group with LDL between 130 and 160. At year 3, in the higher LDL has higher survival rate. However, the group with LDL between 130 and 160 has the highest survival rate at year 5.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |

|  |  |
| --- | --- |
|  |  |
| Incidence rate |
| LDL < 100 | 0.047 |
| 100 < LDL <130 | 0.040 |
| 130 <= LDL <160 | 0.030 |
| LDL >= 160 | 0.029 |

 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | N | Events | S(t) | 95% CI |
| LDL< 100 |  |  |  |  |  |
|  | Year 1 | 163 | 3 | 0.9818 | (0.9447, 0.9941) |
|  | Year 3 | 151 | 12 | 0.9091 | (0.8537, 0.9442) |
|  | Year 5 | 133 | 18 | 0.8 | (0.7304, 0.8534) |
| 130< LDL <=100 |  |  |  |  |  |
|  | Year 1 | 225 | 4 | 0.9825 | (0.9539, 0.9934) |
|  | Year 3 | 209 | 16 | 0.9123 | (0.8673, 0.9425) |
|  | Year 5 | 186 | 23 | 0.8114 | (0.7543, 0.8565) |
| 130<= LDL < 160 |  |  |  |  |  |
|  | Year 1 | 221 | 5 | 0.9778 | (0.9474, 0.9907) |
|  | Year 3 | 210 | 11 | 0.9289 | (0.8865, 0.9558) |
|  | Year 5 | 197 | 13 | 0.8711 | (0.8199, 0.9086) |
| LDL >= 160 |  |  |  |  |  |
|  | Year 1 | 116 | 2 | 0.9829 | (0.9334, 0.9957) |
|  | Year 3 | 111 | 5 | 0.9402 | (0.8786, 0.9710) |
|  | Year 5 | 102 | 9 | 0.8632 | (0.7865, 0.9139) |

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). If *HR* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model, this can be effected by the Stata code

gen fithrA = *HR ^ (ldl* – 160)

It could also be computed by creating a centered LDL variable, and then using the Stata predict command

 gen cldl = ldl – 160

stcox cldl

predict fithrA

* Here is first 10 data list for fithrA:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1~5: | 1.203301 | 1.755244 | 1.395319 | 2.081046 | 1.092898 |
| 6~10: | 0.9780366 | 1.547686 | 1.385028 | 1.305385 | 1.447932 |

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled as a continuous logarithmically transformed variable.
	1. Include full description of your methods, appropriate descriptive statistics (you may refer to problem 1, if the descriptive statistics presented there are adequate for this question), and full report of your inferential statistics.
* The descriptive statistics should see the question 1 without log transformation.
* We still used proportional hazard regression with robust standard error after log transformation. We can get hazard ratio 0.4375, and 95%CI: (0.2966, 0.6452) from STATA. We can calculate the hazard ratio of death is 0.5638 times for each doubling LDL ($HR: 0.4375^{log(2)}=0.5638)$. The hazard ratio of death is 43.62% lower for each doubling LDL. With 95% CI, it is not unusual if the true hazard ratio is between 0.4307 and 0.7381 for each doubling LDL. That is, we are not surprised if the true hazard ratio in higher LDL group is from 26.19% lower to 56.93% lower than lower LDL for each doubling LDL. If alpha=0.05, it has statistically significant (p-value < 0.0001).
	1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). If *HR* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model, this can be effected by the Stata code

gen logldl = log(ldl)

stcox logldl

fithrB = *HR ^ (logldl* – log(160))

It could also be computed by creating a centered logarithmically transformed LDL variable, and then using the Stata predict command

 gen clogldl = log(ldl / 160)

stcox clogldl

predict fithrB

* Here is first 10 estimator for fithrB:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1~5: | 1.150784 | 1.703433 | 1.313884 | 2.219141 | 1.066567 |
| 6~10: | 0.9847617 | 1.462715 | 1.304515 | 1.234544 | 1.363061 |

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled quadratically (so include both a term for serum LDL modeled continuously and a term for the square of LDL).
	1. Include full description of your methods, appropriate descriptive statistics (you may refer to problem 1, if the descriptive statistics presented there are adequate for this question), and full report of your inferential statistics. In the inferential statistics, include your conclusion regarding the linearity of the association of serum LDL and the log hazard.
* The descriptive statistics – see Question1.
* To compare hazard of death, we use proportional hazard regression with robust standard error and with adjusted square of LDL. After adjusted square of LDL, our hazard ratio becomes 0.9742 times for 1 unit difference. 95% CI: (0.9557, 0.9931); With 95% confidence, we are not surprised if the true hazard is 0.69% to 4.43% lower in higher LDL group than lower LDL group for 1 unit difference changed. P-value is 0.008, has statistics significant. When we test if LDL and the square of LDL=0, the p-value is 0.0005, we have highly evidence to say this is nonlinear.
	1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). If *HR* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model for the LDL term and *HR2* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model for the squared LDL term, this can be effected by the Stata code

gen fithrC = *HR^((ldl* - 160)) \* *HR2^(ldl^2* - 160^2)

It could also be computed by creating a centered LDL variable, and then using the Stata predict command

 gen cldl = ldl – 160

 gen cldlsqr= cldl ^ 2

stcox cldl cldlsqr

predict fithrC

* first 10 data for fithrC:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1~5: | 1.096584 | 1.776869 | 1.264096 | 2.513978 | 1.032931 |
| 6~10: | 0.9953307 | 1.447796 | 1.253332 | 1.176875 | 1.322331 |

1. Display a graph with the fitted hazard ratios from problems 1 – 3. Comment on any similarities or differences of the fitted values from the three models.
* The following is a scatter plot with fitted hazard ratio and LDL. In the middle range of LDL level, we can find there is no too much difference between each line; and in the higher LDL or lower LDL, we can see the line which is drawn by fitted hazard ratio from question C is more like U-shaped than the line drawn from question A.



**Discussion Sections: January 27 – 31, 2014**

We continue to discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe the approach to the scientific question posed in the documentation file fev.doc.