**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.

**Methods: A Kaplan Meier curve was produced using LDL categories <=70, 70-100, 100-130, 130-160, and >=160, the restricted mean of survival time was produced, and the restricted mean of each LDL category was also produced. Tabulated survival times at 1000, 1826.25, and 200 days were produces by LDL category. Then robust proportional hazards regression analysis was performed to evaluate the association between serum LDL as a continuous variable and all-cause mortality.**

**Inference: From proportional hazards regression analysis, we estimate that for each 1 unit (mg/dL) difference in ldl value, the risk of all-cause mortality decreases by 0.74% (HR=0.9926) in those with higher ldl. The estimate is statistically significant (p= 0.009). A 95% CI suggests that this observation is not unusual if a group that has 1 unit (mg/dL) increase in ldl value might have a lower risk of mortality that was anywhere from 1.29% to 0.18% (0.9871-0.9982). This proves that there is an association between increasing all cause mortality and decreasing ldl values.**



**The above Kaplan-Meier curve shows overall, lower ldl categories had worst survival, although if we examine the lowest LDL category <=70 (blue), initially it appears that it might have a survival benefit, which then becomes worst survival around 1000 days and eventually has the worst survival of all LDL categories.**

**In a sample size of 735 subjects, the restricted mean of survival time was 1974.469 days with a standard error of 16.566 days and a 95% CI of 1942 days to 2006.94 days**

|  |
| --- |
| **ldlcatg | no of subjects mean Std. Err. [95% Conf. Interval]** |
| **-------------+-------------------------------------------------------------** |
|  **ldl belo | 26 1838.019(\*) 74.4872 1692.03 1984.01** |
|  **ldl 70-1 | 148 1950.288(\*) 37.53749 1876.72 2023.86** |
|  **ldl 100- | 229 1952.222(\*) 30.77626 1891.9 2012.54** |
|  **ldl 130- | 219 1993.479(\*) 29.4222 1935.81 2051.15** |
|  **ldl >=16 | 103 2032.092(\*) 34.66926 1964.14 2100.04** |
| **-------------+-------------------------------------------------------------** |
|  **total | 725 1976.588(\*) 16.3869 1944.47 2008.71** |

**The above table showes the restricted mean survival by LDL category, we see that there is an overall decrease in survival rmean estimate with lower LDL values.**

 **Beg. Survivor Std.**

 **Time Total Fail Function Error [95% Conf. Int.]**

**-------------------------------------------------------------------------------**

**ldl below 70**

 **1000 25 2 0.9231 0.0523 0.7260 0.9802**

 **1826.25 18 7 0.6538 0.0933 0.4402 0.8025**

 **2000 5 1 0.5885 0.1044 0.3600 0.7594**

**ldl 70-100**

 **1000 136 13 0.9122 0.0233 0.8535 0.9480**

 **1826.25 124 12 0.8311 0.0308 0.7603 0.8825**

 **2000 39 3 0.7831 0.0411 0.6890 0.8517**

**ldl 100-130**

 **1000 213 17 0.9258 0.0173 0.8833 0.9532**

 **1826.25 188 25 0.8166 0.0256 0.7601 0.8610**

 **2000 62 1 0.8034 0.0284 0.7407 0.8525**

**ldl 130-160**

 **1000 206 14 0.9361 0.0165 0.8944 0.9616**

 **1826.25 190 16 0.8630 0.0232 0.8100 0.9022**

 **2000 69 4 0.8256 0.0292 0.7594 0.8751**

**ldl >=160**

 **1000 100 4 0.9612 0.0190 0.8998 0.9852**

 **1826.25 91 9 0.8738 0.0327 0.7926 0.9247**

 **2000 35 1 0.8488 0.0402 0.7492 0.9112**

**-------------------------------------------------------------------------------**

* 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

**fithrA has been produced.**

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.

**Methods: A Kaplan Meier curve was produced using logLDL categories (<=70, 70-100, 100-130, 130-160, and >=160). See graph below. Please refer to problem 1a) for the restricted mean of survival, the restricted mean of each LDL category, and the tabulated survival times at 1000, 1826.25, and 200 days that were produced by LDL category There is no value in presenting the rmean of the log LDL categories or list of estimates by log LDL group.**

**Robust proportional hazards regression analysis was performed to evaluate the association between log serum LDL as a continuous variable and all-cause mortality.**

**Inference: From proportional hazards regression analysis, we estimate that for each doubling in ldl value, the risk of all-cause mortality decreases by 43.62% (HR=0.5638) in those with higher ldl. The estimate is highly statistically significant (p= 0.000). A 95% CI suggests that this observation is not unusual if a group that has an LDL value that is twice as high as another group might have a lower risk of mortality that was anywhere from 56.93% to 26.19% (0.4307-0.7381) decrease in mortality. This proves that there is an association between increased all cause mortality and decreasing ldl values.**



**The KM survival curve of log LDL by category is below, it is essentially the same as that of the KM survival curve of LDL, and shows that lower LDL values had worst survival.**

* 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

**fithrB has been produced.**

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.

**Methods: A Kaplan Meier curve was produced using LDL squared categories (<=70, 70-100, 100-130, 130-160, and >=160). See graph below. Please refer to problem 1a) for the restricted mean of survival, the restricted mean of each LDL category, and the tabulated survival times at 1000, 1826.25, and 200 days that were produced by LDL category. Robust proportional hazards regression analysis was performed to evaluate the association between LDL and squared serum LDL as continuous variable and all-cause mortality.**

**Inference: From proportional hazards regression analysis, with a quadratic model with ldl and square of ldl, the HR of ldl is 0.97423 and that of squared ldl is 1.0001, we estimate that for each squaring in ldl value, the risk of all-cause mortality decreases by 0.01% in those with higher ldl. The estimate however is not statistically significant (p= 0.055). A 95% CI suggests that this observation is not unusual if a group that has an LDL value that is squared as high as another group might have a lower risk of mortality that was anywhere from 0.01% to 0.02% (0.9999-1.0002) decrease in mortality. The association of ldl and log hazard is linear with slope of -0.0261 and p value of 0.008, while that of log hazard and sqldl is not linear.**



**The KM survival curve of log LDL by category is below, it is essentially the same as that of the KM survival curve of LDL, and shows that lower LDL values had worst survival.**

* 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

**fithrC produced.**

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.

(?)



**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.