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.

The proportional regression based on Kaplan-Meier estimate is used for association between LDL and mortality. The obstime variable is used for obtaining the time to event or time to censoring depending on which one comes first. The death variable is used for describing event in the Kaplan-Meier estimation. We have 735 subjects in total. Number of failure is 131. From proportional hazards regression, we estimate that for each 1mg/dL unit different in LDL, the risk of death is 99.25% of that in the group with higher LDL. This estimate is highly statistically significant (P < 0.05). A 95% CI suggestion that this observation is not unusual if a group that has a 1 mg/dL higher LDL might have risk of death that was anywhere from 98.7% to 99.8% of that in the group with lower LDL.

* 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

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 proportional regression on log scale based on Kaplan-Meier estimate is used for association between LDL and mortality. The obstime variable is used for obtaining the time to event or time to censoring depending on which one comes first. The death variable is used for describing event in the Kaplan-Meier estimation. We have 735 subjects in total. Number of failure is 131. From proportional hazards regression, we estimate that for each doubling in LDL, the risk of death is 56.3% of that in the group with higher LDL. This estimate is highly statistically significant (P < 0.001). A 95% CI suggestion that this observation is not unusual if a group that has doubling LDL might have risk of death that was anywhere from 43.1% to 73.8% of that in the group with lower LDL.

* 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

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 polynomial regression after putting ldl and squared ldl into the regression model is used for association between LDL and mortality. We used robust option assuming the unequal variance between groups. We have 735 subjects in total. From polynomial regression, we can get the polynomial function as y = -0.00586 \* x + 0.0000181 \* x ^2 + 0.611. We obtain the p-value by test both of ldl and ldlsqr equal to zero. This gives us a highly statistically p-value (P = 0.0105) and indicates there is a non-linear association between LDL and mortality.

* 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

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.

All three different curves share the same trend, that is as the LDL increases to around 160 mg/dL, the HR relative to LDL of 160 mg/dL decreases and this decrease becomes slower and slower. After 160 mg/dL, the relative hazard increases again. However, Proportion HR relative to LDL of 160 mg/dL is much similar to log HR than squared HR. This similarity pattern reflects that the first two share the same principle for obtaining the HR and it is much more reasonable. In practice, we should choose log transformed LDL for doing the regression.

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.