**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:* The survival distribution was estimated using Kaplan-Meier estimates. The hazard of death over the entire period of observation were compared between subjects who differed in serum LDL using proportional hazard regression model. Statistical inference on the hazard of death as a function of serum LDL modeled as a continuous variable was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using Huber-white sandwich estimator. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for proportional hazard regression parameter estimates.**
* ***Results:* Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). There were 151 subjects that died within the observation period (the minimum observation time was 68 days and the maximum 2022 days). The minimum time of follow-up among censored observations was 1,827 days and the maximum 2159 days. From a proportional hazards regression analysis, we estimate that for each 1 mg/dL unit difference in serum LDL, the hazard of death is 0.74% lower in the group with the higher LDL. Based on a 95% confidence interval, this observed ratio of hazard of death suggesting lower hazard of death for groups of patients with higher LDL levels would not be judged unusual if a group that has a 1 mg/dL higher LDL might have hazard of death that was anywhere from 0.18% lower to 1.29% lower than the group with the lower LDL (95% CI for mortality hazard ratio 0.9871 to 0.9982). A two-sided p value of 0.009 suggests that we can with high confidence reject the null hypothesis that the hazard of death over the entire period of observation is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.**
	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.
* ***Methods:* The survival distribution was estimated using Kaplan-Meier estimates. The hazard of death over the entire period of observation were compared between subjects who differed in serum LDL using proportional hazard regression model. Statistical inference on the hazard of death as a function of serum LDL modeled as a continuous logarithmically transformed variable was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using Huber-white sandwich estimator. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for proportional hazard regression parameter estimates.**

***Results:* Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). There were 151 subjects that died within the observation period (the minimum observation time was 68 days and the maximum 2022 days). The minimum time of follow-up among censored observations was 1,827 days and the maximum 2159 days. From a proportional hazards regression analysis, we estimate that for each 2.7183 (e) - fold difference in in serum LDL, the hazard of death is 56.25% lower in the group with the higher LDL. Based on a 95% confidence interval, this observed ratio of hazard of death suggesting lower hazard of death for groups of patients with higher LDL levels would not be judged unusual if a group that has a 2.7183 (e) - fold higher LDL might have hazard of death that was anywhere from 35.47% lower to 70.33% lower than the group with the lower LDL (95% CI for mortality hazard ratio 0.2967 to 0.6453). A two-sided p value of <0.0001 suggests that we can with high confidence reject the null hypothesis that the hazard of death over the entire period of observation is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.**5/5 for performing an appropriate analysis

5/5 for reporting the association appropriately

* Total: 10

	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.
* ***Methods:* The survival distribution was estimated using Kaplan-Meier estimates. The hazard of death over the entire period of observation were compared between subjects who differed in serum LDL using proportional hazard regression model. Statistical inference on the hazard of death as a function of serum LDL modeled quadratically was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using Huber-white sandwich estimator. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for proportional hazard regression parameter estimates.**
* ***Results:* Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). There were 151 subjects that died within the observation period (the minimum observation time was 68 days and the maximum 2022 days). The minimum time of follow-up among censored observations was 1,827 days and the maximum 2159 days. From a proportional hazards regression analysis adjusted second order relationship, we estimate that for each 1 mg/dL unit difference in serum LDL, the hazard of death is 2.58% lower in the group with the higher LDL. Based on a 95% confidence interval, this observed ratio of hazard of death suggesting lower hazard of death for groups of patients with higher LDL levels would not be judged unusual if a group that has a 1 mg/dL higher LDL might have hazard of death that was anywhere from 0.69% lower to 4.43% lower than the group with the lower LDL (95% CI for mortality hazard ratio 0.9557 to 0.9931). A two-sided p value of 0.0005 suggests that we can with high confidence reject the null hypothesis that the hazard of death over the entire period of observation is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.**

**After adjusting for second order relationship, the p value is 0.008, which suggests that we can with high confidence reject the null hypothesis that there is no linearity of the association of serum LDL and the log hazard. And the p-value for after adjusting the linear relationship is 0.055, therefore, we cannot with high confidence to reject the null hypothesis that there is no U-shape trend of the association of serum LDL and the log hazard.**

* 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.

**When the** **population with a defined serum LDL value that is close to 160 mg/dL, the values of fithrA, fithrB, and fithrC are very close to each other, especially when LDL levels are in the range of 90-170 mg/dL. However, when the LDL levels go beyond this range (either smaller or larger), the value of fithrC is markedly larger than fithrB, which is larger than fithrA.**

**The differences come from the different models that have been used. For fithrA, it is a linear model, therefore, the value for firthrA across groups with different LDL level is linear, with lower hazard of death for groups with higher LDL levels.**

**Compared to firthrC, firthrB is more approximating to fithrA. But in the groups with much lower LDL value, the relative hazard ratios rise quickly (compared to fithrA). That is because in 2nd model, serum LDL is modeled as a continuous logarithmically transformed variable.**

**As for firthrC, serum LDL is modeled quadratically. Therefore, a term for the square of LDL is added to the regression model. (As we can see there is an extra element in the equation:** \* *HR2^(ldl^2* - 160^2)**). Therefore, in both end of LDL level (either smaller or larger than 160), the relative hazard ratios rise remarkably.**



**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.