***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 Cox Proportional Hazards regression was performed using a conitinuous measure of serum LDL levels as the predictor variable and instantaneous risk of death (for any cause, over the whole study period) as the response variable. A Wald test was performed to obtain 95% confidence intervals and a p-value. A classical standard error was used that assumes all observations are independent, and equal variance across groups. It was also assumed that the parameters are approximately normally distributed.

Results:

Descriptive Statistics:





10 values were missing from the dataset and were not included in the analysis. Those remaining were split into stratifications of LDL<100 mg/dL, LDL>160 mg/dL, and those with LDL in between 100 and 160 mg/dL. The lowest group contains 174 subjects, the middle group contains 457 subjects, and the highest group contains 107 subjects. The plot of the fitted survival curve for the Cox Proportional Hazards Regression reveals that proportion surviving is higher for those with serum LDL greater than 160 mg/dL during all stages of the data collection (from 0 to 2159 days). The other two stratifications yield more similar survival probabilities: those with serum LDL less than 100 mg/dL have higher survival probabilities from approximately day 0 to day 1000, but for the remainder of the study period those in the mid range having serum LDL between 100 and 160 mg/dL have the higher survival probability of the two strata. Overall, survival is high, with survival probabilities never dropping below 0.7 during the study period.

Inferential Statistics:

According to a Cox Proportional Hazards regression using a continuous measure of serum LDL to estimate instantaneous risk of death, a group with serum LDL 1 unit higher (1 mg/dL) has an instantaneous death rate 0.007 times (0.7%) lower than those with higher LDL, and a group with LDL 10 mg/dL higher has an instantaneous death rate 6.8% lower than those with higher LDL. Based on a 95% confidence interval, this 1-unit multiplicative difference in instantaneous death rate would not be judged unusual if the true 1-unit multiplicative difference were anywhere between 0.013 (1.3%) and 0.002 (0.2%). A two-sided p value of 0.006 suggests statistical significance at the 0.05 alpha level allowing for the rejection of the null hypothesis that the hazard ratio between the two groups has a null value of 1.

* 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:

A Cox Proportional Hazards regression was performed using the logarithmic transformation of a conitinuous measure of serum LDL levels as the predictor variable and instantaneous risk of death (for any cause, over the whole study period) as the response variable. A Wald test was performed to obtain 95% confidence intervals and a p-value. A classical standard error was used that assumes all observations are independent, and equal variance across groups. It was also assumed that the parameters are approximately normally distributed.

Results:

Descriptive Statistics:





10 values were missing from the dataset and were not included in the analysis. Those remaining were split into stratifications of LDL<100 mg/dL, LDL>160 mg/dL, and those with LDL in between 100 and 160 mg/dL. The lowest group contains 174 subjects, the middle group contains 457 subjects, and the highest group contains 107 subjects. The plot of the fitted survival curve for the Cox Proportional Hazards Regression reveals that proportion surviving is higher for those with serum LDL greater than 160 mg/dL during all stages of the data collection (from 0 to 2159 days). The other two stratifications yield more similar survival probabilities: those with serum LDL less than 100 mg/dL have higher survival probabilities from approximately day 0 to day 1000, but for the remainder of the study period those in the mid range having serum LDL between 100 and 160 mg/dL have the higher survival probability of the two strata. Overall, survival is high, with survival probabilities never dropping below 0.7 during the study period.

Inferential Statistics:

According to a Cox Proportional Hazards regression using a log-transformed continuous measure of serum LDL to estimate instantaneous risk of death, for each doubling of serum LDL, the instantaneous death rate is 0.564 times (56.4%) lower than those with higher LDL. Based on a 95% confidence interval, this would not be judged unusual if the true doubling decrease in risk of death were anywhere between 0.407 (40.7%) and 0.775 (77.5%). A two-sided p value of 0.0055 suggests statistical significance at the 0.05 alpha level allowing for the rejection of the null hypothesis that the hazard ratio between the two groups has a null value of 1.

* 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:

A Cox Proportional Hazards regression was performed using a conitinuous measure of serum LDL levels treated quadratically as the predictor variable and instantaneous risk of death (for any cause, over the whole study period) as the response variable. A Wald test was performed to obtain 95% confidence intervals and a p-value. A classical standard error was used that assumes all observations are independent, and equal variance across groups. It was also assumed that the parameters are approximately normally distributed.

Results:

Descriptive Statistics:





10 values were missing from the dataset and were not included in the analysis. Those remaining were split into stratifications of LDL<100 mg/dL, LDL>160 mg/dL, and those with LDL in between 100 and 160 mg/dL. The lowest group contains 174 subjects, the middle group contains 457 subjects, and the highest group contains 107 subjects. The plot of the fitted survival curve for the Cox Proportional Hazards Regression reveals that proportion surviving is higher for those with serum LDL greater than 160 mg/dL during all stages of the data collection (from 0 to 2159 days). The other two stratifications yield more similar survival probabilities: those with serum LDL less than 100 mg/dL have higher survival probabilities from approximately day 0 to day 1300, but for the remainder of the study period those in the mid range having serum LDL between 100 and 160 mg/dL have the higher survival probability of the two strata. Overall, survival is high, with survival probabilities never dropping below 0.7 during the study period.

Inferential Statistics:

The coefficient values from the polynomial Cox Proportional Hazards Regression of 1.00 (exponentiated coefficient for LDL^2) and 0.9261 (exponentiated coefficient for LDL) are difficult to interpret. However, when each term was individually assessed using Wald-based estimates, the p-values are significant to an alpha level of 0.05 (p=0.038 for LDL^2, p=0.00386 for LDL). Therefore, we cannot overlook the significance of the linear term (LDL, with its p=0.00386), and cannot make any conclusions about non-linearity of the relationship between serum LDL and instantaneous risk of death.

* 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.*



 The hazard ratio for problems 1-3 are given by: HR=0.993^(LDL) [1], HR = 0.438^(log(LDL)) [2], and HR = 1.000^(LDL^2)\*0.974^(LDL) [3]. The coefficients were estimated using Cox Proportional Hazards regressions relative to a group having serum LDL of 160 mg/dL, treating LDL as either a generic continuous variable, a log-transformed variable, or a quadratic variable. The difference in each regression’s estimated hazard ratio is due to the various treatments of the predictor variable LDL and thus the construction of each model.

***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.*