**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #4**

February 2, 2015

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 9, 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 investigates associations between death from any cause and age, sex, and serum bilirubin in a population of patients with primary biliary cirrhosis who were enrolled in a randomized clinical trial (RCT) of D-penicillamine. The data can be found on the class web page (follow the link to Datasets) in the file labeled liver.txt. Documentation is in the file liver.doc.

1. Provide suitable descriptive statistics pertinent to the scientific questions addressed in this homework.

**Methods:** Descriptive statistics for age, sex, and bilirubin levels included the sample size and either the mean, standard deviation, minimum, maximum, median and interquartile range (25% and 75%) or the proportion of males & females. Descriptive statistics for the censoring distribution included the minimum and maximum observed censoring times, the mean duration of follow-up time. To summarize survival probabilities by (untransformed) bilirubin level, bilirubin was categorized as follows: [0-1) mg/dL, [1-5) mg/dL, [5-10) mg/dL, [10-20) mg/dL, and [20-30) mg/dL. Within these bilirubin categories, Kaplan-Meier estimates of survival were calculated and graphed.

**Descriptive Statistics:** There were 418 participants who were followed for death from any cause for a Kaplan-Meier average of 6.88 years (median 6.48 years, range 3.33 to 13.13 years), during which 161 deaths were observed. Bilirubin levels were available for all participants at enrollment, and the mean bilirubin was 3.22 mg/dL (SD 4.41 mg/dL, range 0.3 to 28 mg/dL). Age in years was available for all participants at enrollment, and the mean age was 50.74 years (SD 10.45 years, range 26.28 to 74.44 years). Information on sex was available for the 312 participants who were enrolled in the RCT of d-pencillamine and 36 of those individuals were male (11.45%), while 276 were female (88.46%). It is not known whether the participants selected for the RCT were a random sample of total group of 418, or whether the distribution of sex is the same in the larger group.

A graph of the Kaplan-Meier survival curves is shown in Figure 1. The largest difference in survival distributions is seen when comparing the group with the highest bilirubin levels to those with the lowest bilirubin levels. The group with the highest bilirubin levels has the lowest probability of survival.

1. In prior homeworks using the Cardiovascular Health Study datasets, we were able to use logistic regression to investigate associations between mortality and various covariates. Why might such an approach not seem advisable with these data? (Consider the extent to which such analyses might be confounded and/or lack precision.)

Because the data is censored and there are a wide range of observation times, logistic regression is not an accurate method to assess the association between mortality and bilirubin levels. In addition, there may be confounding by gender because biliary cirrhosis is more common in women and there are many more women in the dataset. It is not clear whether the progression of the disease varies by gender, making women more likely to progress to severe liver failure or die from it.

1. Perform a statistical regression analysis evaluating an association between serum bilirubin and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum bilirubin modeled as a continuous variable.
	1. Include a full report of your inference about the association.

**Methods**: Distributions of time to death from any cause was compared across groups defined by baseline bilirubin level using proportional hazards regression with bilirubin as a continuous untransformed variable. The association between bilirubin and all cause mortality was summarized by the hazards ratio, with 95% confidence intervals and two-sided p values.

**Results**: Among the 418 participants, the mean bilirubin was 3.22 mg/dL (SD 4.41 mg/dL, range 0.3 to 28 mg/dL). During an average of 6.88 years of observation, 161 participants were observed to die. From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is a relative 15.2% higher (hazard ratio 1.1524) for each 1 mg/dL increase in baseline bilirubin. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true hazard ratio were anywhere from 12.1% to 18.5% higher in those with a baseline bilirubin level 1 mg/dL higher than the level in another group (95% CI for hazard ratio 1.1208 to 1.1847). Given the two-sided p value of <.0001, we can reject the null hypothesis that the risk of death from any cause is not associated with bilirubin levels and support the finding of higher mortality with higher bilirubin levels.

* 1. For each population defined by serum bilirubin value, compute the hazard ratio relative to a group having serum bilirubin of 1 mg/dL. (This will be used in problem 6). 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 ^ (bili* – 1)

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

 gen cbili = bili – 1

stcox cbili

predict fithrA

See Question 6 for answer.

1. Perform a statistical regression analysis evaluating an association between serum bilirubin and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum bilirubin modeled as a continuous logarithmically transformed variable.
	1. Why might this analysis be preferred *a priori?*

Because biologically, it makes sense that the value of bilirubin in the bloodstream would increase multiplicatively, rather than linearly, with worsening liver disease.

* 1. Include a full report of your inference about the association.

**Methods**: Distributions of time to death from any cause was compared across groups defined by baseline bilirubin level using proportional hazards regression with bilirubin as a continuous logarithmically transformed variable. The association between all cause mortality and bilirubin level was summarized by the hazards ratio, with 95% confidence intervals and two-sided p values

**Results**: Among the 418 participants, the mean bilirubin was 3.22 mg/dL (SD 4.41 mg/dL, range 0.3 to 28 mg/dL). During an average of 6.88 years of observation, 161 participants were observed to die. From a proportional hazards regression analysis, we estimate that the instantaneous risk of death is a relative 98.4% higher (hazard ratio 1.9844) for each 2-fold increase in baseline bilirubin level. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true hazard ration were anywhere from 78% to 121% higher in a group having baseline bilirubin 2-fold higher than that in another group (95% CI 1.78 to 2.21). Given the two-sided p value of <.0001, we can reject the null hypothesis that the risk of death from any cause is not associated with bilirubin levels and support the finding of higher mortality with higher bilirubin levels.

* 1. For each population defined by serum bilirubin value, compute the hazard ratio relative to a group having serum LDL of 1 mg/dL. (This will be used in problem 6). 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 logbili = log(bili)

stcox logbili

fithrB = *HR ^ (logbili*)

(Note that the log(1) = 0 when using any base, so there is no need to rescale by the bilirubin values. Note also that you might want to use a different base in your logarithmic transformation in order to facilitate more natural reporting of effects.)

See answer in problem 6.

1. One approach to testing to see whether an association between the response and the predictor of interest is adequately modeled by an untransformed continuous variable is to add some other transformation to the model and see if that added covariate provides statistically significant improved “fit” of the data. In this case, we could test for “linearity” of the bilirubin association with the log hazard ratio by including both the untransformed and log transformed bilirubin. (Other alternatives might have been bilirubin and bilirubin squared, but in this case our *a priori* interest in the log bilirubin might drive us to the specified analysis.)
	1. Provide full inference related to the question of whether the association is linear.
	2. Again, save the fitted values from this model by obtaining the estimated HRs relative to a group with bilirubin of 1 mg/dl. (This will be used in problem 6.)
2. Display a graph with the fitted hazard ratios from problems 3 - 5. Comment on any similarities or differences of the fitted values from the three models.

Figure 2: Plots of the fitted hazard ratios from proportional hazards regression models including a linear continuous term for bilirubin and a logarithmically transformed term for bilirubin.

1. We are interested in considering analyses of the association between all cause mortality and serum bilirubin after adjustment for age and sex.
	1. What evidence is present in the data that would make you think that either sex or age might have confounded the association between death and bilirubin? (In real life, we would ideally decide whether to adjust for potential confounding in our pre-specified statistical analysis plan (SAP)).
	2. What evidence is present in the data that would make you think that either sex or age might have added precision to the analysis of the association between death and bilirubin? (In real life, we would ideally decide whether to adjust in our pre-specified SAP).
	3. Provide full inference regarding an association between death and bilirubin after adjustment for sex and age.
2. Note that in the above analyses, we completely ignored the intervention in the RCT? What impact could this have had on our results?

Assuming that all participants were randomly assigned to either placebo or treatment in the RCT, the intervention should affect all subgroups of subjects equally. However, if the treatment works better for women or for those with a more severe stage of cirrhosis, then that would skew our results for survival.