**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.

**Method**: I performed a descriptive statistics on age and sex by bilirubin levels. For this, serum bilirubin was categorized into 0, 1, 2, 4, 8 and 16mg/dl. For each bilirubin level, the mean age, its standard deviations, and minimum and maximum values were determined. For gender, percent male and female were determined. And, a Kaplan Meier curve was generated for each bilirubin level to look at survival probabilities at different bilirubin concentrations.

**Result**: The analysis shows that 106 out of 418 patients were missing data on sex. As per the available data, number of patients is predominantly female. The mean age of patients across bilirubin groups are similar, and the percent dead is higher at higher bilirubin level, the mean of which is 3.22mg/dl with standard deviation of 4.41mg/dl and range of 0.3 to 38mg/dl.

From the Kaplan-Meier curve, it is seen that survival probability is low for patients with higher bilirubin levels and it decreases with increasing number of years.

|  |  |
| --- | --- |
|  | **Bilirubin Levels in mg/dl (n = 418)** |
|  | **0-1 (n = 142)** | **1-2 (n = 107)** |
|  | **Mean**  | **Std. Dev.** | **Range** | **Mean**  | **Std. Dev.** | **Range** |
| **Age (years)** | 51 | 9.80 | 31-75 | 51 | 10.79 | 26-77 |
| **Female (%)** | 68.31 |  |  |  | 63.55 |  |
| **Dead (%)** | 85.92 |  |  |  | 71.96 |  |
|  | **2-4 (n = 78)** | **4-8 (n = 48)** |
|  | **Mean**  | **Std. Dev.** | **Range** | **Mean**  | **Std. Dev.** | **Range** |
| **Age (years)** | 50 | 11.35 | 30-72 | 51 | 11.64 | 31-78 |
| **Sex (%)** | 66.67 |  |  | 62.5 |  |  |
| **Dead (%)** | 43.59 |  |  | 33.33 |  |  |
|  | **8-16 (n = 27)** | **16-28 (n = 16)** |
|  | **Mean**  | **Std. Dev.** | **Range** | **Mean**  | **Std. Dev.** | **Range** |
| **Age (years)** | 51 | 10.14 | 33-71 | 50 | 6.37 | 42-66 |
| **Sex (%)** | 59.26 |  |  | 81.25 |  |  |
| **Observation Time (years)** | 25.93 |  |  | 6.25 |  |  |



Figure . Kaplan-Meier based estimates for observation time from study enrollment to death from any cause for 418 patients

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.)

**Answer**: Such an approach is not advisable with these data, because the data is right censored. The minimum period of time for which information on all patients is available is 1.46years only. After this time we have limited and missing data. Since in the Cardiovascular Health Study dataset we had information on all participants, we could perform logistic regression; but since that is not the case here, we can’t do logistic regression.

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.

**Method**: To evaluate an association between serum bilirubin and all-cause mortality by comparing the instantaneous risk of death over the entire period of observation across groups, I stset obstime and status, and performed proportional hazard regression using the Stata command, stcox, followed by predictor variable, bili. Here, the serum bilirubin was treated as a continuous untransformed variable. The confidence interval and two-sided p-values were computed using the Wald’s statistics. Missing data were excluded from the analysis.

**Result**: Data is available on 418 patients among whom 161 patients were observed to have died during an average observation period of 5.25years. Using the proportional hazard regression, the instantaneous risk of death is 1.15 higher for each 1mg/dl increase in serum bilirubin at baseline. Based on a 95% confidence interval, this hazard ratio is not unusual if the true instantaneous risk of death were anywhere between 1.12 and 1.19 higher in a group having baseline serum bilirubin of 1mg/dl. A two sided p-value of <0.001 suggests that we can with high confidence reject the null hypothesis that risk of death from any cause is not associated with serum bilirubin levels in favor for an alternative hypothesis that there is a tendency for higher deaths with higher serum 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

**Answer**: See answer to question 6.

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?*

**Answer**: This analysis is preferred a priori, because serum bilirubin follows a multiplicative scale – the instantaneous risk of death for doubling of bilirubin is not the same as the instantaneous risk of death for four-fold increase in serum bilirubin levels, and so on. When evaluating an association between bilirubin and instantaneous risk of death, logarithmically transformed bilirubin will give a more precise estimate than untransformed bilirubin.

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

**Method**: To evaluate an association between serum bilirubin and all-cause mortality by comparing the instantaneous risk of death over the entire period of observation across groups, I stset obstime and status, and performed proportional hazard regression using the Stata command, stcox, followed by log-transformed predictor variable, logbili. The confidence interval and two-sided p-values were computed using the Wald’s statistics. Missing data were excluded from the analysis.

**Result**: Data is available on 418 patients among whom 161 patients were observed to have died during an average observation period of 5.25years. Using the proportional hazard regression, the instantaneous risk of death is 98% times higher (Hazard Ratio (HR) = 1.984) for each doubling in log-transformed serum bilirubin at baseline. Based on a 95% confidence interval, this hazard ratio is not unusual if the true instantaneous risk of death is anywhere between 1.781 and 2.210 higher. A two sided p-value of <0.001 suggests that we can with high confidence reject the null hypothesis that the risk of health from any cause is not associated with serum bilirubin levels in favor for an alternative hypothesis that there is a tendency for higher deaths with higher serum 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.)

**Answer**: See answer to question 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.

**Method**: To evaluate an association between serum bilirubin and all-cause mortality by comparing the instantaneous risk of death over the entire period of observation across groups, I stset obstime and status, and performed proportional hazard regression using the Stata command, stcox, followed by predictor variables, bili and log-transformed bili. The confidence interval and two-sided p-values were computed using the Wald’s statistics. Missing data were excluded from the analysis.

**Result**: Data is available on 418 patients among whom 161 patients were observed to have died during an average observation period of 5.25years. A proportional hazard regression analysis, modeling both untransformed and log-transformed bilirubin terms, gives a statistically significant association between instantaneous risk of death and bilirubin at baseline (P<0.001). The test of linearity based on untransformed bilirubin in the proportional hazard regression is not statistically significant (P = 0.191), but it is statistically significant for log-transformed bilirubin (P<0.001), which suggests that true association between death from any cause and serum bilirubin is adequately described by log hazard function that is linear in log-transformed bilirubin term.

* 1. 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.)

**Answer**: See answer to question 6.

1. 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 . Plots of the Fitted Hazard Ratios from Three Different Proportional Hazards Regression Models containing: i) a Linear Continuous Serum Bilirubin, ii) Logarithmically Transformed Serum Bilirubin, and iii) Both Linear and Logarithmically Transformed Serum Bilirubin.

**Answer**: As seen in Figure 2, the best linear fit for the relative hazard is given by the model using log transformed data and then by the model using both untransformed and log-transformed data. On the other hand, the untransformed bilirubin does not show a linear relation with relative hazard. It fits similarly to two other models at the lowest data points, but diverges a lot as bilirubin concentration increases, eventually curving upwards near the maximum bilirubin concentration (28mg/dl) observed for our sample. Contrary to this, the fit of both untransformed and log-transformed bilirubin starts out very similar to log-transformed bilirubin fit, but at the higher concentration of the serum bilirubin, starts curving down.

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)).

**Answer**: The age and serum bilirubin scatterplot shows heteroscadasticity. Most of the sample data seems concentrated at the lower concentration end of bilirubin, and variation fans out. There does not seem to be a unidirectional trend in age and bilirubin concentration, but we know that with higher age, there is higher mortality. Thinking in this term, age could be a precision variable. If any association is seen between age and bilirubin at lower concentrations, it could just be a spurious association due to small sample size. On the other hand, being female seems to be strongly related to bilirubin concentrations, but again it could just be an artifact of sampling and is not a confounder as more women with the condition were sampled than men.

* 1. 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).

**Answer**: Being female is associated with the outcome, death from all cause, and age is also associated with the outcome death from all cause. This makes me think that they are precision variables.

* 1. Provide full inference regarding an association between death and bilirubin after adjustment for sex and age.

**Method**: To evaluate an association between serum bilirubin and all-cause mortality by comparing the instantaneous risk of death over the entire period of observation across groups, I performed cox proportional hazards regression with robust standard errors adjusting for age and sex. The model included log-transformed bilirubin, age and sex. The confidence interval and two-sided p-values were computed using the Wald’s statistics. Missing data were excluded from the analysis.

**Result**: Data is available on 312 patients among whom 125 patients were observed to have died during the study period. From the proportional hazard regression analysis model, we see that males have lower risk for mortality for every doubling of serum bilirubin than females keeping all other variables constant, but it is not statistically significant (P<0.799). With increasing age, there is increased instantaneous risk of mortality for every doubling of serum bilirubin, and it is statistically significant (P<0.001).

1. Note that in the above analyses, we completely ignored the intervention in the RCT? What impact could this have had on our results?

**Answer**: Ignoring the intervention in the RCT could have introduced a confounding by treatment in association between serum bilirubin concentration and mortality.