**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #4**

February 2, 2015

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

**Ans: Method: The scientific questions of this homework mainly focus on the association between all-cause mortality and serum bilirubin. Only a part of the variables of the "liver.txt" dataset are of interest, which include three continuous variables (age, serum bilirubin level, observation time), and 3 binary variables (sex, censored status, treatment assignment). Among these variables of interest, age and sex are known to associate with the concentration of serum bilirubin, and also casually associated with mortality. Hence age and sex are potential confounder/precision variable. As this dataset is from a randomized clinical trial, intervention may also modify the association between all-cause mortality and serum bilirubin, hence may be a potential effect modifier.**

**To have a quick glance of the data, we present the number of observations, mean, SD, min, max for continuous variables, age and serum bilirubin level, present percentage for binary variables sex, censored status and treatment assignment. Although observation time is a continuous variable, but we are interested in its quartiles for understanding the censoring distribution. So min, 25%ile, median, 75%ile, max are presented for observation time.**

**In addition, we stratify the data on serum bilirubin levels, with intervals chosen to represent approximate doubling of serum bilirubin levels. In each stratum, Kaplan-Meier curve, Kaplan-Meier estimates of the 10th, 50th, and 90th percentiles and restricted mean are also represented.**

**Results: In the dataset, the total 418 subjects are followed for all-cause death for observation time ranging from 0.11 to 13.14 years. There're 106 missing data of sex at the study enrollment, and no missing data of age, serum bilirubin concentration. Among the 418 subjects with available serum bilirubin levels at enrollment, the mean serum bilirubin is 3.22 mg/dL (SD 4.41 mg/dL). Comparing the KM curves of each stratum, the general trend is that subjects with higher bilirubin level tend to live shorter. The 5-Year Survival Probability is highest in the group with bilirubin concentration lower than 0.50 mg/dL, and lowest in the group with bilirubin concentration higher than 16.00 mg/dL. On average, the subjects with bilirubin concentration between 0.50-0.99 mg/dL have the highest restrict mean of survival, 10.89 years, while the subjects with bilirubin concentration higher than 16.00 mg/dL are estimated to live 2.21 years with the available follow-up data.**

Table1. Summary statistics of variables of interest in the dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Age (years)1 | Serum Bilirubin (mg/dL) 1 | Female (%)3 | Uncensored (%) | Observation time (year) 2 |
| Number of Observations | 418 | 418 | 312 | 418 | 418 |
| Summary Statistics | 50.7(10.4;26.3-78.4) | 3.22(4.41;0.3-28) | 88.50% | 38.50% | 0.11;2.99;4.74;7.16;13.14 |

1 Descriptive statistics is presented as mean (SD; min-max);

2 Descriptive statistics is presented as min;25%ile;median;75%ile;max;

3 106 missing data of sex is omitted for percentage computation.



Figure1. Kaplan-Meier curve for stratum defined by serum bilirubin concentrations

Table 2. Kaplan-Meier estimates for stratum defined by serum bilirubin concentrations

|  |  |
| --- | --- |
|  | Serum Bilirubin at Study Enrollment |
|  | 0.25-0.49mg/dL | 0.50-0.99mg/dL | 1.00-1.99mg/dL | 2.00-3.99mg/dL | 4.00-7.99mg/dL | 8.00-15.99mg/dL | >16.00mg/dL | All level |
| N subjects | 13 | 129 | 107 | 78 | 48 | 27 | 16 | 418 |
| N deaths | 3 | 17 | 30 | 44 | 32 | 20 | 15 | 161 |
| 10%ile of Survival (year)1 | 5.63 | 6.09 | 3.32 | 1.51 | 0.89 | 0.19 | 0.21 | 1.67 |
| Median of Survival (year) 1 | NA | NA | 11.48 | 5.70 | 3.70 | 2.16 | 2.34 | 9.30 |
| 90%ile of Survival (year) 1 | NA | NA | NA | 9.44 | 10.52 | NA | 3.87 | NA |
| 5-Year Survival Probability (%)1 | 100.00 | 92.37 | 87.11 | 52.50 | 32.95 | 30.86 | 6.25 | 70.29 |
| 10-Year Survival Probability (%)1 | 58.33 | 78.04 | 53.85 | 9.52 | 12.20 | NA | NA | 44.22 |
| Restricted Mean of Survival (year) | 9.162 | 10.892 | 9.6152 | 6.0732 | 4.74 | 3.252 | 2.212 | 8.362 |

1The descriptive statistics is based on Kaplan-Meier estimates. NA indicates that the corresponding statistic is not estimable with the available data.

2 largest observed analysis time is censored; mean is underestimated.

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.)

**Ans: In the Cardiovascular Health Study dataset, the minimal observation time of censored data is just above 4 years. And in the previous analysis, we were only interested in the mortality within 4 years. But in the PBC dataset, we're interested in all-cause mortality over all the follow-up time. Doing logistic regression using the data before the minimal observation time cannot answer the scientific question. Also the minimal observation time of censored data is 1.46 years. And the number of subjects who die before 1.46 years after the study enrollment is only 37. The reduction of sample size from 418 to 37 greatly reduces the analysis's precision. (SE may increase more than 3 times.)**

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.

**Ans: Method: There's no missing data of the variables of interest. Hazards of death are compared across groups defined by serum bilirubin using proportional hazards regression (robust) modeling serum bilirubin modeled as a continuous variable. Hazard ratio is summarized to describe the association of all-cause mortality and serum bilirubin level. Confidence interval and two-sided p value is computed using Wald statistics based on the Huber-White sandwich estimator.**

**Results: The total number of subjects is 418 and total number of observed deaths is 161. From proportional hazards regression analysis with robust standard error estimates, we estimate that for each 1 mg/dL unit difference in serum bilirubin, the risk of death is 15.24% higher in the group with the higher serum bilirubin. This estimate is highly statistically significant (P < 0.001). A 95% CI suggests that this observation is not unusual if a group that has a 1 mg/dL higher bilirubin might have risk of death that was anywhere from 12.08% higher to 18.47% higher than the group with the lower serum bilirubin.**

* 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

**Ans: Fitted values are to be displayed in problem 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?*

**Ans: A constant difference in serum bilirubin level would not be expected to lead to same increase in risk. Instead, we expect same increase in risk for multiplicative changes in serum bilirubin level. So this analysis is the preferred *a priori.***

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

**Ans: Method: There's no missing data of the variables of interest. Hazards of death are compared across groups defined by serum bilirubin using proportional hazards regression (robust) modeling serum bilirubin modeled as a log (base 2) transformed continuous variable. Hazard ratio is summarized to describe the association of all-cause mortality and serum bilirubin level. Confidence interval and two-sided p value is computed using Wald statistics based on the Huber-White sandwich estimator.**

**Results: The total number of subjects is 418 and total number of observed deaths is 161. From proportional hazards regression analysis with robust standard error estimates, we estimate that for each doubling in serum bilirubin, the risk of death is 98.45% higher (hazard ratio = 1.984) in the group with the higher serum bilirubin. This estimate is highly statistically significant (P < 0.001). A 95% CI suggests that this observation is not unusual if a group that has a bilirubin twice as high as another might have risk of death that was anywhere from 1.781 to 2.212 times as high as the group with the lower serum bilirubin.**

* 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.)

**Ans: Fitted values are to be displayed 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.

**Ans: Method: There's no missing data of the variables of interest. Hazards of death are compared across groups defined by serum bilirubin using proportional hazards regression (robust) modeling serum bilirubin as a combination of a continuous untransformed serum bilirubin and a log (base 2) transformed continuous serum bilirubin concentration. Associations between all-cause mortality with linear bilirubin and log (base 2) transformed bilirubin are tested simultaneously tested. Two-sided p value and CI are computed from Wald statistics with the Huber-White sandwich estimator. In addition, a test for nonlinearity of the association between hazard and serum bilirubin is performed using a Wald test with the null hypothesis that the coefficient of the log (base 2) transformed bilirubin is zero.**

**Results: The total number of subjects is 418 and total number of observed deaths is 161. From proportional hazards regression analysis (robust) modeling bilirubin concentration as a combination of a continuous untransformed serum bilirubin concentration and a log (base 2) transformed continuous serum bilirubin concentration, there's statistically significant evidence (P < 0.001) for association between hazard of all-cause death and serum bilirubin at the study enrollment. P value of the test for nonlinearity based on log (base 2) transformed serum bilirubin is smaller than 0.001, hence suggests that we can with high confidence reject the null hypothesis that the association between all-cause mortality and serum bilirubin is adequately modeled by an untransformed continuous serum bilirubin.**

* 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.)

**Ans: Fitted values are to be displayed in problem 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.

**Ans: Fitted hazard ratios are plotted in figure 2. Proportional hazard regression modeling on the combination (problem 5) shows much similarity with the log fit model (problem 4), especially when the bilirubin concentration is smaller than 15 mg/dL. These two models diverge with bilirubin concentrations higher than 18 mg/dL, where the data points are relatively sparse. But either of these two models is obviously different from the linear fit model (problem 3). Both the comparison and the nonlinearity test for the association between hazard and serum bilirubin in problem 5 shows that the association is not adequately modeled by an untransformed continuous serum bilirubin. Also we expect the hazard of all-cause death to act on the multiplicative scale of serum bilirubin. The model of problem 4, proportional hazard regression modeling on the log (base 2) transformed serum bilirubin, is appropriate in this setting.**



Figure 2. Fitted hazard ratios from proportional regressions modeling on untransformed continuous serum bilirubin, modeling on log (base 2) transformed serum bilirubin and modeling on a combination of untransformed and log (base 2) transformed serum bilirubin respectively. The relative hazard ratios are compared with groups with bilirubin of 1mg/dL.

1. We are interested in considering analyses of the association between all cause mortality and serum bilirubin after adjustment for age and sex.

Table 3. Summary Statistics of Age and Sex stratified on serum bilirubin at the Study Enrollment

|  |  |
| --- | --- |
|  | Serum Bilirubin at Study Enrollment |
|  | 0.25-0.49mg/dL | 0.50-0.99mg/dL | 1.00-1.99mg/dL | 2.00-3.99mg/dL | 4.00-7.99mg/dL | 8.00-15.99mg/dL | >16.00mg/dL | All level |
| Age1 (year) | 53.7(7.06; 43.0-69.0);n=13 | 50.9(10.01; 30.6-75.0);n=129 | 50.6(10.79; 26.3-76.7);n=107 | 50.2(11.35; 29.6-71.9);n=78 | 50.9(11.64; 30.9-78.4);n=48 | 50.8(10.14; 33.2- 70.6);n=27 | 50.2(6.37; 41.9-65.8);n=16 | 50.7(10.45; 26.3-78.4);n=418 |
| Female (%)2 | 100;n=11 | 92.5;n=93 | 88.3;n=77 | 82.5;n=63 | 78.9;n=38 | 94.1;n=17 | 100;n=13 | 88.5;n=312 |

* 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)).

**Ans: (Sex) The percentage of females differs across the groups defined by the serum bilirubin at study enrollment, which means that sex is associated with serum bilirubin level in the sample. Also sex is believed to be causally associated with all-cause mortality, which is clear in figure 3. So Sex should be adjusted, as it's potential confounder.**

**The association of sex and serum bilirubin is clearer if we compare the distributions of serum bilirubin in males and females. In the dataset, for females the mean bilirubin concentration is 3.31 mg/dL with standard deviation of 4.75 mg/dL (median concentration is 1.30 mg/dL), while for males the mean bilirubin concentration is 2.87 mg/dL with standard deviation of 2.23 mg/dL (median concentration is 2.20 mg/dL).**

**The coefficient of 0.4339 of linear model of bilirubin on sex is also an evidence of that sex is associated with serum bilirubin concentration.**

****

**Figure 3. Kaplan-Meier estimates across groups defined by sex (106 missing data of sex are omitted)**

* 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).

**Ans: (Age) Across the groups defined by the serum bilirubin at the study enrollment, the distribution of age is quite similar. Though the group with lowest bilirubin level has slightly high mean age and smaller variance, its small sample size and the small difference cannot convince us that it may confound the analysis.**

**The coefficient of 0.0001 from linear regression modeling bilirubin on age is also an evidence of that age is not associated with serum bilirubin concentration in this sample.**

**But age is known to be casually associated with all-cause mortality, which is obvious in figure 4.**

**So age is a precision variable. Adjusting age may add precision to the analysis, as controlling age decreases the within-group variance of response.**

**(Sex) As stated in problem 7a, we think sex is a confounder, whether adjusting sex would add precision depends on the equation** .

**But the equation is hard to evaluate in this case.**

****

**Figure 4. Kaplan-Meier estimates across groups defined by age**

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

**Ans: Method: 106 missing data of sex are omitted from the analysis. Hazards of death are compared across groups defined by log (base 2) transformed continuous serum bilirubin, using proportional hazards regression (robust). Continuous age and binary sex are also modeled as the regression covariate for assess the association between death and bilirubin after adjustment for age and sex. Two-sided p value and CI are computed from Wald statistics with the Huber-White sandwich estimator.**

**Results: The total number of subjects in this analysis is 312 and total number of observed deaths is 125. From proportional hazards regression analysis (robust) modeling the log (base 2) transformed continuous bilirubin, we estimate that for each doubling in serum bilirubin, the hazard is 2.108 times as high as the group with lower serum bilirubin. This estimate is highly statistically significant (P < 0.001). A 95% CI suggests that this observation is not unusual if a group that has a bilirubin twice as high as another might have risk of death that was anywhere from 1.840 to 2.416 times as high as the group with the lower serum bilirubin.**

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

**Ans: Intervention in the RCT is conceptually independent with the serum bilirubin concentration, as the subjects with different bilirubin should be randomized to receive treatment or control in the RCT. But it's still possible that we have different distributions of serum bilirubin level across the intervention arms. (Actually the mean bilirubin in treatment arm is 2.881 mg/dL with SD 3.639 mg/dL, while the mean bilirubin in the placebo arm is 3.667 mg/dL with SD 5.293 mg/dL.)**

**Also intervention is conceptually associated with the response, all-cause death, in a causal way.**

**Intervention in the RCT could have confounded our results.**

**But in the actual situation, both of treatment or placebo in the RCT have very similar Kaplan Meier curve (figure 5). And from the proportional hazard regression, the estimated hazard ratio comparing treatment arm to placebo arms is 0.9448 with a two-sided p value 0.752. It suggests that the intervention is not associated with all-cause mortality. Hence, I believe that intervention in the RCT acts as a variation inflator in our analysis.**

****

**Figure 5. Kaplan-Meier estimates across groups defined by intervention in the RCT**