**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.

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.

**Answer:**

**Methods:** The scientific question here is to investigate the associations between death from any cause and age, sex and serum bilirubin among patients with primary cirrhosis in an RCT. Death from any cause is our outcome, we treat death as our event(outcome) of interest, represented by two variables: Obstime-observation time (days) and Status-survival status(0=censored). Thus only 5 variables were used here to address this scientific question: namely age bili obstime sex status. And due to the fact that this is a RCT, this study design prevents potential confounders ( measured or unmeasured) at baseline(right after randomization). Age, sex and serum bilirubin may be potential effect modifier or precision variable to the associations of each other and the outcome, depending on which variable is our main predictor of interest. Also because this is an RCT among patients with primary biliary cirrhosis, I think serum bilirubin should be our main predictor of interest based on our knowledge that serum bilirubin level is more evidently and biologically an indicator of liver function. Then, age and sex are potential effect modifiers or precision variables to the association between serum bilirubin and death.

First, to have a basic idea of data, for continuous variables( age, serum bilirubin and observation time) , mean, SD, min, max were presented. And for binary variables(sex, survival status), percentages were used.

To describe the association between serum bilirubin and survival , scatterplots are of no scientific meaning here to describe censored data as we have here. So we choose KM curves across strata to do this job. But first we need to categorize the continuous variable-serum bilirubin. Based on prior experience , a constant difference in serum bilirubin would not indicate the same increase in risks for our outcome(death), I would expect a multiplicative effect on the risk of death from serum bilirubin level. So, I stratified serum bilirubin levels with intervals for a 5 fold of serum bilirubin. I generated a new variable called biliCAT5a, for serum bilirubin level [0.3-1.5)mg/dl, biliCAT5a=1, for serum bilirubin level [1.5-7.5)mg/dl, biliCAT5a=2,for serum bilirubin level [7.5-28]mg/dl, biliCAT5a=3.

**Results:** Descriptive statistics were shown in Table 1. Data is available on 418 participants’ age, serum bilirubin level , survival status and observation time. Only 312 out of 418 participants had information on sex. The mean observation time was 1918 days, with the minimum of 41 days and the maximum of 4795 days. The mean serum bilirubin level among our participants was 3.22(mg/dL), with a standard deviation of 4.41 and a range from 0.30 to 28.00 (mg/dL).

From Figure1 we can see: generally patients in biliCTG51=1, which is [0.3-1.5)mg/dl had the best survival during the entire follow-up. The KM curve of this groups stayed constantly above the other two groups’, and there was no overlapping. The KM curve of patients with the serum bilirubin [1.5-7.5)mg/dl stayed in the middle, and patients with the serum bilirubin [7.5-28]mg/dl had the worst survival comparing to the other two groups. Its KM curve stayed constantly below the other two.

**Table1**. Descriptive statistics of participants’ age, sex, serum bilirubin and survival

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age (years)1 | Serum Bilirubin (mg/dL) 1 | Sex2 | Uncensored (%) | Censored(%) | Observation time (days)1 |
| Female (%) | Male(%) |
| Number of Observations | 418 | 418 | 275 | 37 | 163 | 254 | 418 |
| Descriptive Statistics | 50.74(10.45;26.28-78.44) | 3.22(4.41;0.30-28.00) | 88.00% | 22% | 39.00% | 61% | 1918(1105;41.00-4795) |

1 **Descriptive statistics: mean(standard deviation;minimum-maximum)**

2 **106 participants having missing values for sex**

 **Figure1**. Kaplan-Meier curve for strata defined by serum bilirubin levels1



1**For serum bilirubin level [0.3-1.5)mg/dl, biliCAT5a=1, for serum bilirubin level [1.5-7.5)mg/dl, biliCAT5a=2,for serum bilirubin level [7.5-28]mg/dl, biliCAT5a=3**

I also calculated the estimates of surviving probability at specific times (here I chose 365th , 1825th , 2920th days, corresponding to 1st , 5th and 8th year). For patients with serum bilirubin level [0.3-1.5)mg/dl, the survival probabilities at these three time points were 0.98, 0.90 and 0.77 respectively. for patients with serum bilirubin level [1.5-7.5)mg/dl, the survival probabilities at these three time points were 0.85, 0.43 and 0.28 respectively.For those with serum bilirubin level [7.5-28]mg/dl, the survival probabilities at these three time points were 0.70, 0.21 and 0.08 respectively. These results agreed with the KM curves in Figure1, suggesting a better survival for patients with lower serum bilirubin levels in the follow-up time.

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:** Because for the Cardiovascular Health Study, the minimal observed time of censoring data is just 1480 days(a little bit more than 4 years). Then we dichotomized the time to death according to death within 4 years of study enrollment or death after 4 years. Our outcome became a binary variable-death within 4 years or not. So we could use logistic regression to conduct the analyses relevant to our scientific question. However, in this dataset, we are investigating the associations between death from any cause and age, sex and serum bilirubin . We are not fixed at one point of time, instead we want to comparing the survival situation across groups during the entire follow up time. Our outcome here is not a binary variable anymore. So our scientific question doesn’t allow for a logistic regression.

Even if we changed our scientific question to investigating the association between death after or before the minimum censored time and age, sex or serum bilirubin. Our minimum censored time here is 533 days, with only 36 deaths happened before 533th day. This would be too little events and would not make much scientific sense to investigate the survival of an approximately 1.46 years (533 days) follow up for a chronic disease.

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.

**Answer:**

**Methods:** Participants with missing values in either serum bilirubin , observation time or survival status were excluded from analysis. Hazards of death over the entire follow up time are compared across groups defined by serum bilirubin using proportional hazards regression with Robust standard error. Serum bilirubin level is modeled as a continuous variable. Hazard ratio is summarized to evaluate the association between 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:** 418 patients had full data on serum bilirubin , observation time and survival status. From proportional hazards regression analysis, we estimate that for each 1 mg/dl unit difference in serum bilirubin , the risk of all-cause death is 15.24% higher in the group with 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 12.09% 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

**Answer:** See 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?*

**Answer:** Based on our knowledge, a constant difference in serum bilirubin level would not be expected to confer the same increase in risk. A multiplicative effect on risk may be better, as we would expect the same increase in risk for each doubling of serum bilirubin. Thus we use log transformed serum bilirubin to fit the model.

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

**Answer:**

**Methods:** Participants with missing values in either serum bilirubin , observation time or survival status were excluded from analysis. Hazards of death over the entire follow up time are compared across groups defined by log transformed (base=2) serum bilirubin using proportional hazards regression with Robust standard error. Log transformed (base=2) serum bilirubin level is modeled as a continuous variable. Hazard ratio is summarized to evaluate the association between 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:** 418 patients had full data on serum bilirubin , observation time and survival status. From proportional hazards regression analysis, we estimate that for each doubling in serum bilirubin , the risk of all-cause death is 98.45% higher in the group with 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 serum bilirubin twice as high as another might have risk of all cause death that was anywhere from 1.78 to 2.21 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.)

**Answer:** See 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.

**Answer:**

**Methods:** Participants with missing values in either serum bilirubin , observation time or survival status were excluded from analysis. Hazards of death over the entire follow up time are compared across groups defined by log transformed (base=2) bilirubin and untransformed serum bilirubin using proportional hazards regression with Robust standard error. Log transformed (base=2) and untransformed serum bilirubin level are modeled as a continuous variables. Hazard ratio is summarized to evaluate the association between 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:** 418 patients had full data on serum bilirubin , observation time and survival status. From proportional hazards regression analysis, the point estimate for log transformed (base=2) serum bilirubin is 2.27, with a confidence interval of 1.83-2.82, and a statistically significant P value (P<0.001). While the point estimate for untransformed bilirubin is 0.96, with a confidence interval of -1.45-1.01, and a statistically non-significant P value of 0.148.This suggests that by including both Log transformed (base=2) and untransformed serum bilirubin, log transformed (base=2) serum bilirubin could sufficiently capture most of the association between all-cause death and serum bilirubin, hence we can with high confidence reject the null hypothesis for the “linearity” of the bilirubin association with the log hazard ratio or the association between all-cause death 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.)

**Answer:** See 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.

**Figure 2**. Fitted hazard ratios from proportional regressions modeling on “untransformed serum bilirubin”, on “ log (base 2) transformed serum bilirubin” and on a both “untransformed and log (base 2) transformed serum bilirubin”. 

**Answer:**

The first order trends are all upwards, indicating that the hazard ratio for all cause deaths increases on average with higher level of serum bilirubin. However, the red line (modeled on the log transformed（base 2)) has the best linear fit, it is more straighter than the other two lines. This suggests by log transforming our serum bilirubin, it adequately captured a linear trend (on average) between hazard ratio for all cause deaths and serum bilirubin. The 1st line (modeled only with untransformed serum bilirubin), is the least linear one, indicating maybe untransformed serum bilirubin is not a good measurement to investigate the linear association between hazard ratio for all cause deaths and serum bilirubin. The combination of log transformed and untransformed serum bilirubin gives us a line in the middle of the two lines discussed above in terms of linearity, giving further evidence that including untransformed serum bilirubin in the model is not a good idea.

This line in return supports our decision in problem 4 choosing a log transformed（base 2) serum bilirubin in the model.

Thus, a constant difference in serum bilirubin level would not be expected to confer the same increase in risk. A multiplicative effect on risk is better, as we would expect the same increase in risk for each doubling of serum bilirubin. Thus we use log transformed serum bilirubin to better fit the model more.

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:**

 As shown in Table 2 below, we can see that in this dataset, the mean age of patients varies according to different strata of serum bilirubin. And the percentage of female patients also differs by serum bilirubin strata. This tells us that age and sex are associated with our POI(predictor of interest: serum bilirubin) in this sample.

 Also, based on our common knowledge, age and sex are obviously causally associated with the outcome here(all cause deaths). With the same serum bilirubin level, the survival would very possibly be different for a man and a women, or for a 40-year old and a 60-year old.

Thus sex and age could be the potential confounders for the association of our scientific question.

**Table 2.** Descriptive Statistics of Age and Sex by on serum bilirubin category

|  |  |
| --- | --- |
|  | Serum Bilirubin (mg/dl) |
|  | [0.3-1.5) | [1.5-7.5) | ≥7.5 |
| Age1 (year) | 50.98(10.33; 26.28-76.71);n=223 | 50.9(10.01; 30.6-75.0);n=152 | 50.6(10.79; 26.3-76.7);n=43 |
| Female2 | 92.77%;n=166 | 50.44%;n=116 | 50.56%;n=30 |

 1 **Descriptive statistics: mean(standard deviation;minimum-maximum)**

 2 **percentage of female**

* 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:**

Comparing the standard deviations of age in Table 1(10.45) and the standard deviations of age across different strata (10.33, 10.01 and 10.79), there are not very much difference between these standard deviations. So I don’t think adjust for age would increase our precision of analysis.

Speaking of sex, only 312 out of 418 patients having information on their sex, there is a too much high proportion on the missing values of sex. I highly doubt it would increase our precision when adjust for sex in our PH regression model.

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

**Answer:**

**Methods:** Participants with missing values in either serum bilirubin , observation time or survival status were excluded from analysis. Hazards of death over the entire follow up time are compared across groups defined by log transformed (base=2) bilirubin and age and sex using proportional hazards regression with Robust standard error. Hazard ratio is summarized to evaluate the association between 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:** 312 patients had full data on serum bilirubin , observation time, survival status, sex and age. From proportional hazards regression analysis, we estimate that for each doubling in serum bilirubin , the risk of all-cause death is 2.11 times as high comparing to the group with lower serum bilirubin but the same sex and of the same age. This estimate is highly statistically significant (P<0.001). A 95% CI suggests that this observation is not unusual if a group with the same age and sex and has a serum bilirubin twice as high as another might have risk of all cause death that was anywhere from 1.83 to 2.42 times as high as the group with the lower serum bilirubin but same sex and age.

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:** Randomization would remove the associations between our POI(serum bilirubin) and the potential confounders(measured and unmeasured). So age and sex would not be confounders to the association between all cause deaths and serum bilirubin. They would only be precision variables now, as sex and age are only associated with our outcomes here in this sample. Adjusting sex and age would increase our precision in accessing the association between all cause mortality and serum bilirubin.