**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:** For descriptive statistics of the survival probabilities by serum bilirubin level, subjects are divided into four groups based on their serum bilirubin levels: less than 2 mg/dL, 2-4 mg/dL, 4-8 mg/dL, and greater than or equal to 8 mg/dL. Descriptive statistics are presented in the table below of each group as well as the entire sample set. Kaplan-Meier curves of these groups were graphed and estimates of survival were calculated for the 2, 5 and 10 year survival probabilities. For continuous variable serum bilirubin level, the mean, standard deviation and the data range are included.

**Results:** The dataset contains 418 subjects and there are no missing data in their serum bilirubin level. Among these 418 subjects, the mean serum bilirubin level was 3.22 mg/dL (SD 4.41 mg/dL, range 0.3 to 28 mg/dL). To study the association between the all-cause mortality and serum bilirubin level, subjects are divided into four groups based on their serum bilirubin level: less than 2 mg/dL, 2-4 mg/dL, 4-8 mg/dL, and greater than or equal to 8 mg/dL. These subjects were followed for death from any cause for 13 years and 161 deaths were observed during this period. Table 1 presents estimates of the 2, 5 and 10 year survival probabilities for each groups as well as the entire sample set. It is apparent that the survival probability decreases with the increase of the serum bilirubin level. For example, the 5 year survival probability is lowest (20.56%) in the group with their bilirubin level greater than or equal to 8 mg/dL and is observed highest in the group with their bilirubin level less than 2mg/dL (90.46%). Figure 1 presents the Kaplan-Meier survival probability. It is also clear from the figure that the survival distributions are different among these groups and the group with the highest bilirubin level has the lowest survival probability.

Table 1: Kaplan-Meier based estimates of survival probability for subjects with different serum bilirubin measurements at baseline.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Serum Bilirubin level | | | | Total  n=418 |
|  | <2(mg/dL)  (n=249) | 2-4(mg/dL)  (n=78) | 4-8(mg/dL)  (n=48) | ≥8(mg/dL)  (n=43) |
| 2 year Survival Probability | 0.9597 | 0.8462 | 0.7703 | 0.6047 | 0.8802 |
| 5 Year Survival Probability | 0.9046 | 0.525 | 0.3295 | 0.2056 | 0.7029 |
| 10 Year Survival Probability | 0.6702 | 0.0952 | 0.122 | N/A | 0.4422 |



Figure 1: Kaplan-Meier based estimates of distribution of time for subjects with different serum bilirubin measurements at baseline.

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:** In the previous Cardiovascular Health Study datasets, there is a much bigger sample size (n =3879) and the first censored data appeared at 4 years. Therefore, when dichotomizing the data according to death within or after 4 years, there are still enough data for us to calculate the odd ratio and do logistic regression. However, in this data set, there are only 418 subjects and the first censored subject appeared early in the study at 1.45 year. Therefore, we cannot dichotomize the data. Also since the sample pool is small, the variance would be larger and the final logistic regression result would lack precision.

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:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum bilirubin and all-cause mortality. Distributions of time to death from any cause were modeled on the serum bilirubin level as a continuous untransformed variable. The slop of the regression model was obtained to estimate the hazards ratio and its 95% confidence interval (CI) was also determined based on Wald statistics.

**Results:** There are a total of 418 subjects in the study with a mean serum bilirubin level of 3.22 mg/dL (SD 4.41 mg/dL; range 0.3 – 28 mg/dL). From proportional hazards regression analysis, we estimate that for each 1 mg/dL difference in the serum bilirubin level, the risk of death is 15.2% higher in the group with the higher serum bilirubin. With 95% confidence, it would not be unusual if the true instantaneous risk of death were anywhere between 12.1% and 18.5% higher for the group with higher serum bilirubin compared to the group with 1mg/dl lower serum bilirubin. This result is highly statistically significant (P < 0.001). Therefore, we reject the null hypothesis that the all-cause mortality is not associated with the serum bilirubin level and conclude that the group with higher bilirubin is associated with higher risk of death.

* 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

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:** First of all, as a serum protein, bilirubin is likely to act on a multiplicative scale. Therefore, it would be more reasonable to use a multiplicative model to assess the association.

Also, in this sample set, the level of serum bilirubin is different by orders of magnitude between subjects. Therefore, it would be better to work with log transformed bilirubin levels, which can downweight outliers.

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

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum bilirubin and all-cause mortality. Distributions of time to death from any cause were modeled on the serum bilirubin level as a continuous logarithmically transformed variable. The slop of the regression model was obtained to estimate the hazards ratio and its 95% confidence interval (CI) was also determined based on Wald statistics.

**Results:** There are a total of 418 subjects in the study with a mean serum bilirubin level of 3.22 mg/dL (SD 4.41 mg/dL; range 0.3 – 28 mg/dL). From proportional hazards regression analysis, we estimate that for each 10% increase in the serum bilirubin level, the risk of death is 9.88% higher in the group with the higher serum bilirubin. With 95% confidence, it would not be unusual if the true instantaneous risk of death were anywhere between 8.26% and 11.5% higher for the group with 10 % higher serum bilirubin compared to the group with lower serum bilirubin. This result is highly statistically significant (P < 0.001). Therefore, we reject the null hypothesis that the all-cause mortality is not associated with the serum bilirubin level and conclude that the group with higher bilirubin is associated with higher risk of death.

* 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.)

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:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum bilirubin and all-cause mortality. Distributions of time to death from any cause were modeled on the serum bilirubin level as a continuous logarithmically transformed variable as well as a logarithmically transformed bilirubin level. The coefficients for the untransformed bilirubin term and the log transformed bilirubin term were obtained in the regression model and its 95% confidence interval (CI) was also computed based on Wald statistics.

**Results:** Based on the result of a proportional hazards regression analysis, the p-value of the test is less than 0.0001 considering the coefficients of both untransformed bilirubin and log transformed bilirubin. This result is statistically significant at and we reject the null hypothesis that the all-cause mortality is not associated with serum bilirubin level. The p-value of the nonlinearity test for the untransformed bilirubin level is 0.149, indicating that we do not have high confidence that the all-cause mortality is not adequately modeled by the log transformed continuous bilirubin level. The p-value of the test on the nonlinearity for the log bilirubin is < 0.0001, and therefore we have high confidence that the mortality cannot be adequately modeled by the untransformed bilirubin level.

* 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.)

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 above demonstrates the fitted hazard ratios for groups with different serum bilirubin levels obtained when using three proportion hazards regression models. All three models indicate that the hazard ratio is higher for group with higher bilirubin levels. The fitted values are similar between the model based on the logarithmically transformed bilirubin and the model based on both untransformed and log transformed bilirubin. The fitted values from the model based on only the untransformed bilirubin are quite different from the other two. This indicates that the mortality cannot be adequately modeled by only the untransformed bilirubin level. And the bilirubin level is likely to be associated with mortality at a multiplicative scale.

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)).

**Methods:** A linear regression analysis was performed on an association between mean serum bilirubin level and variables (sex, age) using Huber-White estimates of the standard error. The slop of the regression model and the 95% confidence interval were obtained.

**Results:** Based on the linear regression analysis, we estimate that the mean bilirubin level is 0.434mg/dL higher in female compared to male (P=0.353). We also estimate for every 1 year difference in age between two groups of subjects, the mean bilirubin level is 0.001mg/dL in the older group (P=0.952). However, in both cases, the difference in means is small and the P-values are not statically significant. We can not reject the null hypothesis that there is no association in the average bilirubin level across sex and age groups. Therefore, both sex and age are not confounder variables when studying the association between all-cause mortality and the serum bilirubin level.

* 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).

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between all-cause mortality and variables (sex, age). The slop of the regression model was obtained to estimate the hazards ratio and its 95% confidence interval (CI) was also determined based on Wald statistics.

**Results:** From proportional hazards regression analysis, we estimate that the risk of death is 38.4% lower in female compared to male (P=0.043, 95% CI 1.43% to 61.5%). We also estimate for every 1 year difference in age between two groups of subjects, the risk of death is 4.00% higher in the older group (P<0.001, 95% CI 2.47% to 5.54%). Both results are statistically significant. Therefore, we reject the null hypothesis that the all-cause mortality is not associated with sex and age. Combined with the results in 7a, we can not reject the null hypothesis that there is no association in the average bilirubin level across sex and age groups. But both of them are associated with the all-cause mortality. Therefore, sex and age are precision variables when studying the association between all-cause mortality and the serum bilirubin level.

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

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between all-cause mortality and serum bilirubin. Sex and age were considered as precision variables in the regression model. Distributions of time to death from any cause were modeled on the serum bilirubin level as a continuous logarithmically transformed variable. The slop of the regression model was obtained to estimate the hazards ratio and its 95% confidence interval (CI) was also determined based on Wald statistics.

**Results:** There are a total of 418 subjects in the study with a mean serum bilirubin level of 3.22 mg/dL (SD 4.41 mg/dL; range 0.3 – 28 mg/dL). From proportional hazards regression analysis, we estimate that for each 10% increase in the serum bilirubin level for the subjects having the same age and sex, the risk of death is 10.8% higher in the group with the higher serum bilirubin. With 95% confidence, it would not be unusual if the true instantaneous risk of death were anywhere between 8.74% and 12.9% higher for the group with 10% higher serum bilirubin compared to the group with lower serum bilirubin. This result is highly statistically significant (P < 0.001). Therefore, we reject the null hypothesis that the all-cause mortality is not associated with the serum bilirubin level and conclude that the group with higher bilirubin is associated with higher risk of death.

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:** The intervention in this study might affect the all-cause mortality. However it would not affect the serum bilirubin level since it is a randomized trial. Therefore, the intervention might be a precision variable when studying the association between serum bilirubin level and all-cause mortality. Therefore, ignoring the intervention will not affect the point estimate but will affect the CI and fitted values from the model.