**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #4**

January 27, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 3, 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 builds on the analyses performed in homeworks #1, #2, and #3. As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information.

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled as a continuous variable.
	1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.

**METHODS for descriptive statistics:** The table below presents descriptive statistics for the data. The table presents the estimated survival probability using the Kaplien-Mier Estimates at 1,3 and 5 years and we categorized the predictor of interest based on the categorization of serum LDL provided in HW 1 where combine two groups. Effectively we have the group for LDL level ideal and ideal for high risk category with LDL levels from 100mg/dL and below, then we have the category for near ideal to borderline LDL level from 100 to 160 mg/dL and finally we have the high LDL group with LDL from 160 mg/dL and above. We also provide survival probability estimates without adjusting for LDL or the combined estimates. We then provide a plot of the Kaplien Mier (Kaplan Meier) survival curves for the three groups. All the estimates are rounded off to three significant figures.

**DESCRIPTION:** Based on the descriptive statistics below we notice a general trend where the survival probabilities are better for the patients with serum LDL levels greater than or equal to 160 mg/dL. At the end of year 1,3 and 5 we notice that the estimated survival probability is higher for the patients with high serum LDL however, based on the descriptive statistics alone we cannot comment on the statistical or scientific significance. If we consider the plot of the Kaplien Mier survival estimates curve we observe the trend mentioned before where for high LDL we have seemingly better survival curves however, based on the curve wee notice that curves for medium and low LDL intersect at some point which is reason to doubt the assumption about linearity.

The descriptive statistics provided below are based on the data of 725 subjects out of a total 735 patients in the study. Data for serum LDL was missing for 10 subjects and they were not included the analysis for descriptive or inferential statistics.

|  |  |
| --- | --- |
|  | Kaplien-Mier Estimated survival probability |
|  | low LDL <100 mg/dLn=174 | Medium LDL100-160 mg/dLn=448 | High LDL>= 160 mg/dLn=103 | Unadjustedn=725 |
| Year 1 | 0.983 | 0.980 | 1.00 |  0.981 |
| Year 3 | 0.914 | 0.920 | 0.952 | 0.921 |
| Year 5 | 0.8046 | 0.839 | 0.874 | 0.835 |



**METHODS:** We fit a proportional hazards regression model using the right censored data for time to death as our response and serum LDL as our predictor of interest for 725 subjects as the 10 subjects with missing LDL levels were not included. Time to death is modelled as a continuous right censored variable measured in days and serum LDL is modelled as a continuous predictor measured in mg/dL. The estimates were obtained by using maximum partial likelihood estimation. For our model we allowed for the possibility of hetroscedasticity and hence obtaining robust standard error estimates and corresponding confidence intervals. For our analysis we used Wald test Statistics and similarly the confidence intervals were obtained using a normality assumption.

**INFERENCE:** From proportional hazards regression analysis on 725 subjects , we estimate that for each 1 mg/dL unit difference in serum LDL, the risk of death is 0.738% lower in the group with the higher serum LDL i.e. the estimated hazard ratio is 0.993. This estimate is statistically significant (two sided P = 0.009< .05) at a 95% confidence level. A 95% confidence level suggests that this observation is not unusual if a group that has a 1 mg/dL higher serum LDL might have risk of death that was anywhere from 0.182 % lower to 1.29% lower than the group with the lower serum LDL, i.e. the estimated hazard ratio would not be surprising if the true hazard ratio between groups differing by 1mg/dL serum LDL was anywhere from 0.987 to 0.998. Based on a statistically significant results we reject the null hypothesis that the there is no difference in the survival experience for different levels of serum LDL in favor of the alternative that serum LDL is associated with instantaneous risk of death.

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). 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 ^ (ldl* – 160)

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

 gen cldl = ldl – 160

stcox cldl

predict fithrA

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled as a continuous logarithmically transformed variable.
	1. Include full description of your methods, appropriate descriptive statistics (you may refer to problem 1, if the descriptive statistics presented there are adequate for this question), and full report of your inferential statistics.

**METHOD FOR DESCRITIVE STATISTICS:** For this question we refer to question 1 and the descriptive statistics provided there. The reason being that based on the different descriptive statistics we can present for a right censored response variable we can only discuss possible associations in the sample by considering the predictor as a categorical variable. The descriptive statistics at our disposal cannot be used to judge the functional form of the hazard ratio in the sample. As mentioned in class a scatter plot of the predictor versus response would not be reasonable for right censored data. Having said that, we present some descriptive statistics for the predictor of interest which is serum LDL and the log of serum LDL. The reason for doing so would be to see that for the sample if a log transformation could be reasonable for instance we could not justify a log transformation if we had any measurements of 0mg/dL serum LDL.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean  | Standard Deviation | Min  | Max |
| Serum LDL mg/dL | 125.8 | 33.60 | 11.00 | 247.0 |
| Log(LDL) log(mg/dL) | 4.796 | 0.2928 | 2.398 | 2.509 |

 

**DESCRIPTION:** Based on the descriptive statistics presented in the table above we see that since the lowest value for the variable serum LDL is 11 there is no error in the data and we would not run into any issues if we take the log of such a data. The plots above are the histograms to represent the distribution of serum LDL in the data. The histogram on the left is for serum LDL and the histogram on the right is for log(serum LDL). The histogram for serum LDL exhibits the variance of the variable and taking the log of this variable certainly decreased the variance however, one obvious issue with the log transform is that while for ldl we notice some outliers they do not seem that extreme compared to the data but with a log transform we get an obvious and fairly extreme outlier. How this outlier will modify any inference cannot be determined with the given information but this issue is certainly notable and unexpected results should ensure the outliers do not bias the results.

For inference and model fitting we have the following methods and results:

**METHODS:** We fit a proportional hazards regression model using the right censored data for time to death as our response and log of the serum LDL as our predictor of interest for 725 subjects as the 10 subjects with missing LDL levels were not included. Time to death is modelled as a continuous right censored variable measured in days and log of serum LDL is modelled as a continuous predictor measured in mg/dL. The estimates were obtained by using maximum partial likelihood estimation. For our model we allowed for the possibility of hetroscedasticity and hence obtaining robust standard error estimates and corresponding confidence intervals. For our analysis we used Wald test Statistics and similarly the confidence intervals were obtained using a normality assumption. Because we have log transformed data the comparison is no longer for an absolute increase say by 1 mg/dL but is specified as a k-fold increase. For instance in our analysis we talk about a 2-fold increase or the estimates for hazard ratio are reported as a two fold increase. We could have obtained this two fold increase by scaling our predictor by log(2) or by directly evaualuating the estimate from a fitted model which represents an e-fold increase where e is the Euler’s constant approximately 2.71. In my analysis we chose to directly evaluate the hazard ratio for a 2-fold increase in serum LDL using the formula $HRe^{log⁡(2)}$ where HRe represents the e-fold hazard ratio estimate. Similarly the confidence intervals were generated by a similar formula as above only now HRe is replaced by the lower and upper limits of the confidence interval.

**INFERENCE:** From a proportional hazards regression analysis on 725 subjects , we estimate that for doubling our serum LDL, the risk of death is 43.6% lower in the group with the higher serum LDL i.e. the estimated hazard ratio is 0.564. This estimate is highly statistically significant (two sided P = < .0005) at a 95% confidence level. A 95% confidence level suggests that this observation is not unusual if a group that has a higher serum LDL might have risk of death that was anywhere from 26.2 % lower to 56.9% lower than the group with the lower serum LDL, i.e. the estimated hazard ratio would not be surprising if the true hazard ratio between groups was anywhere from 0.431 to 0.738. Based on a statistically significant results we reject the null hypothesis that the there is no difference in the survival experience for different levels of serum LDL in favor of the alternative that serum LDL is associated with instantaneous risk of death.

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). 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 logldl = log(ldl)

stcox logldl

fithrB = *HR ^ (logldl* – log(160))

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

 gen clogldl = log(ldl / 160)

stcox clogldl

predict fithrB

1. Perform a statistical regression analysis evaluating an association between serum LDL and all-cause mortality by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled quadratically (so include both a term for serum LDL modeled continuously and a term for the square of LDL).
	1. Include full description of your methods, appropriate descriptive statistics (you may refer to problem 1, if the descriptive statistics presented there are adequate for this question), and full report of your inferential statistics. In the inferential statistics, include your conclusion regarding the linearity of the association of serum LDL and the log hazard.

Based on the descriptive statistics provided so far and the reasoning behind why many descriptive statistics cannot be evaluated or are meaningless due to the fact that we have right censoring, we refer the reader to the descriptive statistics provided in question 1 and 2.

**METHODS:** We fit a proportional hazards regression model using the right censored data for time to death as our response and the serum LDL and the square of serum LDL as our predictors of interest for 725 subjects as the 10 subjects with missing LDL levels were not included. Time to death is modelled as a continuous right censored variable measured in days and log of serum LDL is modelled as a continuous predictor measured in mg/dL and the square of serum LDL is measured in mg^2/dL^2. The estimates were obtained by using maximum partial likelihood estimation. For our model we allowed for the possibility of hetroscedasticity and hence obtaining robust standard error estimates and corresponding confidence intervals. For our analysis we used Wald test Statistics and similarly the confidence intervals were obtained using a normality assumption. Because we have transformed data using a quadratic transformation the comparison is no longer for an absolute increase say by 1 mg/dL but is difficult to specify and hence is not a question we answer here in our analysis. We do however, consider the assumption of linearity and test for that using a Wald test. Since we are using robust standard errors we cannot use a likelihood ratio test but we feel a Wald test would be sufficient in this case and we also consider the assumption of the variable serum LDL being significant by testing all the parameters in the model using a Wald test.

 **INFERENCE:** From a proportional hazards regression analysis on 725 subjects , we estimate a coefficient of 0.974 (95% CI 0.953,0.996)for serum LDL and for squared serum LDL we estimate a coefficient of 1.00 (95% CI 0.999,1.00). We test for association by testing both serum LDL and squared serum LDL simultaneously which shows that serum LDL significant (two sided P-value = 0.0250 <0.05) which means we reject the null hypothesis that there is no association in favor of the alternative that serum LDL is associated with all cause mortality. We then continue to test for non-linearity and we do not find clear evidence of non-linearity (two sided P-value = 0.0890>0.05).

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). If *HR* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model for the LDL term and *HR2* is the hazard ratio (use the actual hazard ratio estimate) obtained from your regression model for the squared LDL term, this can be effected by the Stata code

gen fithrC = *HR^((ldl* - 160)) \* *HR^((ldl* - 160)^2)

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

 gen cldl = ldl – 160

 gen cldlsqr= cldl ^ 2

stcox cldl cldlsqr

predict fithrC

1. Display a graph with the fitted hazard ratios from problems 1 – 3. Comment on any similarities or differences of the fitted values from the three models.



For the plot above we notice that for LDL levels close to160mg/dL the three curves seem to be almost the same but they diverge out at the ends. We observe a higher difference between the three fits for smaller values of serum LDL and for higher LDL values we notice that the linear and log fit are similar whereas the quadratic fits a slightly higher relative hazard.

As far as the shape of the curve is concerned we do observe a u-shaped curve for the quadratic fit and log decreasing fit for the log transform. Based on these plots it is not possible to judge how the fit is for each transformation of the data but we do notice that for serum LDL around 60 to 190 mg/dL the trend seems to be fairly linear because when we fit a quadratic curve with more degrees of freedom for curve fitting we see that it largely agrees with the linear fit on the that range.

The differences in the curves for the LDL values outside the above mentioned range could also be due to outliers in the data for LDL .

**Discussion Sections: January 27 – 31, 2014**

We continue to discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe the approach to the scientific question posed in the documentation file fev.doc.