**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #4**

January 27, 2014

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: In order to study a potential association between serum LDL and all-cause mortality, we compared the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL. Proportional hazards regression with robust standard errors was used to address this question. Serum LDL was treated continuously and measured in mg/dL No other covariates were considered. Survival from time of enrollment in the study was measured in days, but is presented in years for ease of interpretation. Wald-based estimates are used for point estimates and CI, while robust methods were used to compute standard errors.

For the descriptive purposes, participants were divided into low, medium and high LDL categories as described in the table below. A Kaplan-Meier plot is also included.

Inference: The data available are from a study of 735 elderly Americans from four communities. 10 of these participants do not have LDL levels recorded, and so are excluded from the study.

 From the plot and the table, we see that the survival probabilities are similar for all three categories of LDL (low, medium and high). At year 4 and beyond, those with low LDL (<129mg/dL) have consistently lower survival probabilities, but in the first years all three are quite similar. The medium and high LDL lines cross each other several times, suggesting little difference in the two. The table is color coded to show which of the three categories has the highest (green) and lowest (red) KM survival probability at that time. We see that high LDL has the highest for years 1, 2, and 3. The low LDL group has the lowest probabilities for years 2, 3, 4, and 5.

The Proportional hazards regression model estimate that for each 1mg/dL increase in serum LDL, the risk of death is 0.738% lower in the group with higher LDL (hazard ratio of .9926). This estimate is highly statistically significant (two sided p-value = 0.009). A 95% confidence interval suggests that this observation would not be unusual if a group with 1mg/dL higher LDL risk of death anywhere from 0.182% to 1.29% lower that the group with the lower LDL.



|  |  |
| --- | --- |
|  | *Kaplan-Meier Survival Probabilities* |
| Time | Low LDL level (<129mg/dL) | Medium LDL level (130-159 mg/dL) | High LDL level (>160mg/dL) |
| 1 Year | 0.982 | 0.978 | 1.000 |
| 2 Years | 0.949 | 0.956 | 0.981 |
| 3 Years | 0.911 | 0.929 | 0.953 |
| 4 Years | 0.873 | 0.911 | 0.907 |
| 5 Years | 0.807 | 0.871 | 0.869 |

* 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

Predicted values were computed.

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.

Methods: In order to study a potential association between serum LDL and all-cause mortality, we compared the instantaneous risk (hazard) of death over the entire period of observation across groups defined by log serum LDL. Proportional hazards regression with robust standard errors was used to address this question. The log of serum LDL measurements was calculated, and was treated continuously. No other covariates were considered. Survival from time of enrollment in the study was measured in days, but is presented in years for ease of interpretation. Wald-based estimates are used for point estimates and CI, while robust methods were used to compute standard errors.

Since log preserves order, the descriptive statistics in question 1 apply here as well.

Inference: The same data was used here as in question 1; description of the sample is above. A robust proportional hazards model estimates that for each doubling in serum LDL, the risk of all-cause mortality is .564 times as high in the group with higher LDL. A 95% confidence interval suggests that this observation would not be unusual if a group that has twice as high LDL levels had anywhere from .431 to .738 times as high as the group with lower LDL. With high confidence (two sided-p-value < .001) we reject the hypothesis that log of serum LDL and all-cause mortality are not associated.

* 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

Predicted values were computed.

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.

Methods: In order to study a potential association between serum LDL and all-cause mortality, we compared the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL modeled quadratically. Proportional hazards regression with robust standard errors was used to address this question. Serum LDL and the square of these measurements were the two predictors used. Both were treated continuously. No other covariates were considered. Survival from time of enrollment in the study was measured in days, but is presented in years for ease of interpretation. Wald-based estimates are used for point estimates and CI, while robust methods were used to compute standard errors.

The descriptive statistics in question 1 divide LDL measurements in three groups, allowing to study possible non-linearity of risk of death across groups defined by LDL. The interpretation in part 1 applies here as well. Focusing on the potential non-linearity, we see that survival probabilities for those with medium levels of LDL tend to be in between survival probabilities for those with high and low levels. This suggests a linear relationship would not miss clear trends, at least as far as can be seen from KM descriptive statistics.

Inference: The same data was used here as in question 1; description of the sample is above. A robust proportional hazards model with both a linear and a quadratic term of LDL estimates that for an increase of 1 unit from x to x+1 mg/dL of serum LDL, the risk of all-cause mortality is .9743\*e0.000152x times as high in the group with higher LDL. In this model, as in the other groups with higher LDL are estimated to have lower risk of all-cause mortality. We have little evidence (two-sided p-value = .055) that LDL and probability of death are not associated linearly. In other words, the quadratic term of LDL adds little precision to our estimate of instantaneous risk of all-cause mortality.

* 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)) \* *HR2^(ldl^2* - 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 cldlsqrs

predict fithrC

Predicted values were computed.

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.

Plot:



Comments: As we can see from the plot the predicted values are indeed quite different, especially for extreme values of serum LDL. Although the quadratic term was not significant, there is clearly a difference between C and the linear model A. Model B, which fit log LDL also predicts higher risk for extreme values of 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.