**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: The instantaneous risk of death over the study observation period was evaluated across groups defined by serum LDL using a Cox proportional hazards regression model with standard errors determined by the Huber-White sandwich estimator. Serum LDL was modeled as a continuous variable, and the two-sided p-values and 95% confidence intervals (CI) were Wald-based estimates. From this analysis, the hazard ratio of death according to serum LDL was determined by evaluating the slope. The threshold for statistical significance is 0.05.

Descriptive Statistics: Of 735 randomized patients, 725 had recorded serum LDL measurements. Patients included males and females from age 65 to age 99 years, with a mean age of 74.5 years (standard deviation (SD) 5.45 years), and 49.7% were male. The mean serum LDL was 125.8 mg/dL (SD 33.6 mg/dL; 95% CI 123.4, 128.3 mg/dL; range 11 – 247 mg/dL). During the 5-year study, a total of 119 (16.4%) patients died, and 606 (83.6%) patients survived at least 5 years.

Results: From a simple proportional hazards regression analysis, we estimate the instantaneous risk of death is 0.31% lower for each 1 mg/dL increase in serum LDL. The observed hazard ratio is 0.997 for the incrementally higher LDL, which, based on the 95% CI, would be consistent if the true population hazard ratio were between 0.995 and 0.999. From the logrank test, the two-sided p-value is 0.0064, suggesting there is sufficient evidence to reject the null hypothesis that there is no difference in survival based on serum LDL levels.

* 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.

Methods: The instantaneous risk of death over the study observation period was evaluated across groups defined by serum LDL using a Cox proportional hazards regression model with standard errors determined by the Huber-White sandwich estimator. Serum LDL was logarithmically transformed prior to model fitting to introduce a multiplicative effect on risk. The two-sided p-values and 95% confidence intervals (CI) were Wald-based estimates. The hazard ratio of death according to serum LDL was determined by exponentiation of the slope from the proportional hazards regression analysis to determine the relative risk. The threshold for statistical significance is 0.05.

Descriptive Statistics: See problem 1 above for descriptive statistics describing the sample in this study.

Results: From a Cox proportional hazards regression analysis with logarithmically (assuming natural log) transformed serum LDL values, we estimate the instantaneous risk of death is 56.2% lower for every e-fold increase in serum LDL. The hazard ratio is estimated to be 0.437, which is consistent with a true population hazard ratio between 0.297 and 0.645. The two-sided p-value is <0.0001 suggesting sufficient evidence to reject the null hypothesis of no difference in survival based on serum LDL levels. This is not a clinically useful measure, so instead, we estimate the risk of death is 7.6% lower for every 10% increase in serum LDL. From the 95% CI, the hazard ratio of 0.924 is consistent if the true population hazard ratio were between 0.891 and 0.959. From the logrank test, the two-sided p-value is <0.0001, suggesting there is sufficient evidence to reject the null hypothesis that there is no difference in survival based on serum LDL levels.

* 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.

Methods: The instantaneous risk of death over the study observation period was evaluated across groups defined by serum LDL using a Cox proportional hazards regression model with standard errors determined by the Huber-White sandwich estimator. Serum LDL was quadratically transformed prior to model fitting by adding a squared term to the linear serum LDL analysis. So the model included regression models for LDL and for LDL2 giving a model equation of $log\left(λ\left(X\_{i}\right)\right)=log\left(λ\left(t\right)\right)+β\_{1}×LDL+β\_{2}×\left(LDL\right)^{2}$. The two-sided p-values and 95% confidence intervals (CI) were Wald-based estimates. The hazard ratio of death according to serum LDL was determined by multiplying the slopes from the linear LDL and squared LDL model analyses, since the hazard ratio for each 1 mg/dL increment of LDL is $e^{β\_{1}}×e^{β\_{2}}$. The threshold for statistical significance is 0.05.

Descriptive Statistics: See problem 1 above for descriptive statistics describing the sample in this study.

Results: From a Cox proportional hazards regression analysis with quadratically transformed serum LDL values, we estimate the instantaneous risk of death is less than 0.682% lower for every 1 mg/dL increase in serum LDL. When the terms are evaluated separately, the linear LDL term results in an estimated hazard ratio of 0.9932 (95% CI 0.9809, 1.0056) with a p-value of 0.279, while the squared LDL term results in an estimated hazard ratio of 1.000014 (95% CI 0.99997, 1.000061) with a p-value of 0.547. Neither term is statistically significant, however, to answer the question of whether or not the association is linear, the terms must be tested simultaneously. When analyzed simultaneously, the hazard ratio is estimated to be 0.9932, which is consistent with a true population hazard ratio between 0.981 and 1.006. When analyzed together, the two-sided p-value is 0.0185, suggesting sufficient evidence to reject the null hypothesis of no difference in survival based on serum LDL levels. There is statistically significant evidence that the trend in 5-year all-cause mortality versus mean LDL levels is nonlinear.

* 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^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 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.



Figure 1. Comparison of relative hazard of death according to serum LDL relative to an individual with serum LDL = 160 mg/dL.

The three fitted models use different transformations to analyze the data, but all three showed statistical significance such that we rejected the null hypothesis of no association between serum LDL and 5-year all-cause mortality. From the graph it is clear that the relative hazard for death at a given serum LDL level compared to serum LDL = 160 mg/dL appears to trend downward with increasing LDL for all three models, but the rate of decline and shape of the resultant curve is different between all three, particularly when serum LDL is below approximately 150 mg/dL. At LDL levels greater than about 150 mg/dL there is no obvious difference in relative hazard estimated by each model. The linear model assumes homoscedasticity and a linear trend in the log hazard ratio across groups. By introducing a logarithmic transformation, some possible heteroscedasticity is attenuated and a multiplicative effect on the risk is introduced. Finally, the quadratic transformation adds a 2nd order linear term and is modeled to identify a U-shaped trend in the data. The obvious deviation between the three models in Figure 1 above at the lower levels of serum LDL is consistent with the method of LDL modeling because both logarithmic and quadratic transformations accentuate the relative hazard as values of LDL diverge from 160 mg/dL, although since the quadratic transformation is a modification of linear terms, the result is more similar to a singular linear LDL analysis. Also, both the logarithmically transformed and quadratically transformed result in relative risk estimates that follow a curve as LDL increases, compared to the straight line by untransformed LDL values. Since the transformed data produces p-values <0.05, it suggests the data may be nonlinear.