**4644 Total : 34 points.**

**Biost 515 (Winter 2014)**

**Instructor: Scott Emerson**

**Homework 4**

*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. 10 points.
	1. Include full description of your methods, appropriate descriptive statistics, and full report of your inferential statistics.

**Methods:** The survival distribution was estimated using Kaplan-Meier estimates and modeling serum LDL (mg/dL) as a continuous predictor of interest. A hazard ratio and 95% confidence interval were computed using Cox proportional hazards regression with the Huber-White sandwich estimator of the standard errors. Descriptive statistics are presented by strata defined by serum LDL less than or equal to 129 mg/dL, serum LDL between 130 mg/dL and 159 mg/dL, and serum LDL greater than or equal to 160 mg/dL. Method parts: 3 points.

**Results:** Data is available on 735 subjects, however 10 of those subjects are missing data on serum low density lipoprotein (LDL). Those subjects are omitted from all analyses, but we cannot assess the impact that such omissions might have on the generalizability of our results. The following table and graph depict Kaplan-Meier estimates of survival probability for the 393 subjects whose serum LDL was less than or equal to 129 mg/dL, 225 subjects whose serum LDL was between 130 mg/dL and 159 mg/dL, and 107 subjects whose serum LDL was greater than or equal to 160 mg/dL. From the graph, there is clear evidence that subjects with the highest LDL ( ≥ 160 mg/dL) have a greater likelihood of survival than subjects with the lowest LDL (≤ 129 mg/dL). Subjects with serum LDL between 130 mg/dL and 159 mg/dL slightly overlapped with its neighboring strata. Appropriate descriptive statistic: 4points.

From proportional hazards regression analysis, 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 LDL. This estimate is statistically significant (two sided p-value = 0.009) and a 95% confidence interval suggests that this observation is not unusual if a group that has a 1 mg/dL higher LDL might have risk of death that was anywhere from 0.182% to 1.290% lower than the group with the lower LDL. Therefore, we reject the null hypothesis of no association between serum LDL and survival probability in favor of a hypothesis that higher LDL is associated with greater survival probability. Interpretation: 4 points. Scott used 10 mg/dl unit.

|  |
| --- |
| Table 1: Kaplan-Meier Survival Estimates |
|  | Survival Probabilities |
| Time | LDL ≤ 129 mg/dL | LDL 130 – 159 mg/dL | LDL ≥ 160 mg/dL |
| 1 year | 0.982 | 0.978 | 1.000 |
| 2 years | 0.949 | 0.956 | 0.981 |
| 3 years | 0.911 | 0.930 | 0.952 |
| 4 years | 0.873 | 0.908 | 0.913 |
| 5 years | 0.807 | 0.869 | 0.874 |



* 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

 **Answer:** Done.

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. 9 points.

**Methods:** The survival distribution was estimated using Kaplan-Meier estimates and modeling serum LDL (mg/dL) as a continuous logarithmically transformed predictor of interest. Based on prior experience, a log transformation of LDL may result in more precise estimates since unit differences in serum LDL have a multiplicative effect. A hazard ratio and 95% confidence interval were computed using Cox proportional hazards regression with the Huber-White sandwich estimator of the standard errors. Descriptive statistics are presented in problem 1 of this homework. P values and CIs are computed using Wald statistics. Method parts: 2 points.

**Results:** From proportional hazards regression analysis, we estimate that for each doubling in serum LDL (mg/dL), the risk of death is 43.6% lower in the group with the higher LDL. This estimate is highly statistically significant (p-value < 0.001) and a 95% confidence interval suggests that this observation is not unusual if a group that has serum LDL twice as high as another group might have risk of death that was anywhere from 26.2% to 56.9% lower than the group with the lower LDL. Therefore, we reject the null hypothesis of no association between serum LDL and survival probability in favor of a hypothesis that higher LDL is associated with greater survival probability.4 points.

* 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

 **Answer:** Done.

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. 7 points.

**Methods:** The survival distribution was estimated using Kaplan-Meier estimates and modeling serum LDL (mg/dL) quadratically. (That is, we used a second order model to include a term for LDL treated continuously and a term for the square of LDL.) The Wald test facilitates in determining the statistical significance of an association between serum LDL and survival. As it is difficult to quantify a quadratic association, a plot displaying the fitted values from proportional hazards regression is used to interpret the instantaneous risk of death among this population. Descriptive statistics are presented in problem 1 of this homework. 2 points. Missing method part for a second test for the significance of squared term.

**Results:** From proportional hazards regression analysis, modeling serum LDL as a quadratic variable, we can observe in the graph (next page) that the instantaneous risk of death, when LDL is modeled as a quadratic variable, greatly increases as LDL drops below 100 mg/dL. It similarly increases for high levels of LDL, although the risk is not as high. Clearly, the model estimates a U-shaped curve, and this curve takes on a rather obtuse shape. According to the Wald test for association, these estimates are not statistically significant (p-value = 0.055). Therefore, we do not find clear evidence that the trend in survival versus serum LDL is nonlinear, and we fail to reject the null hypothesis of no quadratic association [just put “no association” here] between serum LDL and survival probability. 2 points. Interpretation for Cox regression model results is missing.



* 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

**Answer:** Done.

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.

**Results:** The following graph displays the fitted hazard ratios relative to a group having serum LDL of 160 mg/dL for three statistical regression analyses: (1) modeling LDL as a continuous variable, (2) modeling LDL as a logarithmically transformed continuous variable, and (3) modeling LDL as a quadratic variable. All three models have a U-shaped distribution, meaning that the instantaneous risk of death (relative to subjects with LDL ≥ 160 mg/dL) is greatest when serum LDL is unusually high or low. Note that, again for all three models, the highest instantaneous risk of death is estimated for groups with the lowest serum LDL. As shown, modeling LDL as a continuous variable produces conservative risk estimates for all levels of serum LDL; while modeling LDL as a logarithmically transformed variable increases the estimated risk exponentially as serum LDL levels drop. When LDL is modeled quadratically, the estimated instantaneous risk of death is higher than the other two models for more extreme LDL (outliers excluded) and lower than the other two models for midrange LDL

8 points. It should be noted that the U-shape seen in the quadratic fit cannot be taken as proof that the highest LDL groups actually have increase risk over the groups with moderate levels: A quadratic curve ultimately has to be U-shaped over the whole real line, just as the linear and logarithmic curves must be monotonic (steadily increasing or steadily decreasing).

