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 survival distribution was estimated using Kaplan-Meir estimates. Descriptive statistics were generated for strata defined by serum LDL levels less than 130mg/dL, 130-159mg/dL, and greater than or equal to 160mg/dL. Difference in survival distributions between strata defined by serum LDL as a continuous variable was tested using Cox proportional hazards regression. The hazard ratios and 95% confidence interval was computed using the Huber-White sandwich estimator of the standard errors.**

**Descriptive Statistics:**

****

**(months)**

|  |  |
| --- | --- |
|  | **Kaplan-Meier Survival Probabilities by Serum LDL** |
|  | **≤ 129mg/dL****(n=393)** | **130-159mg/dL****(n=225)** | **≥ 160mg/dL****(n=107)** |
| **1 years** | 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 |

**Inference: The above graph and table depict Kaplan-Meier estimates of survival probability for the 393 subjects with serum LDL less than 130mg/dL, the 225 subjects with LDL between 130mg/dL and 159mg/dL, and the 107 subjects with serum LDL of at least 160mg/dL. The graph and table show a trend for higher survival probabilities in groups with higher serum LDL levels. However, there are several points of crossover between the curves and table values. From proportional hazards regression analysis of 725 subjects, 131 of whom were observed to die during the study, we estimate that for each 10mg/dL unit difference in serum LDL levels the instantaneous risk of death is 7.135% lower in the group having higher LDL. A 95% confidence interval suggests that this observation is not unusual if the true instantaneous risk of death was anywhere from 1.805% and 12.177% lower risk of death in a group having baseline serum LDL 10 mg/dL higher than another group. Using a two-sided p-value this observation is statistically significant at a 0.05 level of significance (p = 0.0093) and we reject the null hypothesis that the instantaneous risk of death is not associated with serum LDL levels in favor of a tendency for lower mortality for higher 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

**The fitted hazard ratio variable was created using the predict function and the stata code listed above.**

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 survival distribution was estimated using Kaplan-Meir estimates. Descriptive statistics were generated for strata defined by serum LDL levels less than 130mg/dL, 130-159mg/dL, and greater than or equal to 160mg/dL. Difference in survival distributions between strata defined by log transformed serum LDL as a continuous variable was tested using Cox proportional hazards regression. The hazard ratios and 95% confidence interval was computed using the Huber-White sandwich estimator of the standard errors.**

**Descriptive Statistics: See problem 1 descriptive statics. The same plot and table were used because the untransformed serum LDL values are the most interpretable.**

**Inference: The graph and table shown in problem 1 depict Kaplan-Meier estimates of survival probability for the 393 subjects with serum LDL less than 130mg/dL, the 225 subjects with LDL between 130mg/dL and 159mg/dL, and the 107 subjects with serum LDL of at least 160mg/dL. The graph and table show a trend for higher survival probabilities in groups with higher serum LDL levels. However, there are several points of crossover between the curves and table values. From proportional hazards regression analysis of 725 subjects, 131 of whom were observed to die during the study, we estimate that for each doubling of serum LDL levels the instantaneous risk of death in the group having higher LDL is 0.564 times that of the group having lower LDL. A 95% confidence interval suggests that this observation is not unusual if the true instantaneous risk of death for the group having higher LDL was anywhere from 0.431 to 0.738 times that of the group having lower LDL. Using a two-sided p-value this observation is statistically significant at a 0.05 level of significance (p < 0.0001) and we reject the null hypothesis that the instantaneous risk of death is not associated with serum LDL levels in favor of a tendency for lower mortality for higher serum LDL levels.**4/5 for performing an appropriate analysis

Did not report which statistic the statistical inference is based on (-1)

4/5 for reporting the association appropriately

Not enough interpretation of point estimate (-0.5)

Not enough interpretation of CI (-0.5)

Total: 8

* 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

**The fitted hazard ratio variable was created using the predict function and the stata code listed above.**

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 survival distribution was estimated using Kaplan-Meir estimates. Descriptive statistics were generated for strata defined by serum LDL levels less than 130mg/dL, 130-159mg/dL, and greater than or equal to 160mg/dL. Difference in survival distributions between strata defined by serum LDL as a continuous variable modeled quadratically was tested using Cox proportional hazards regression. The hazard ratios and 95% confidence interval was computed using the Huber-White sandwich estimator of the standard errors.**

**Descriptive Statistics: See problem 1 descriptive statics. The same plot and table were used because the untransformed serum LDL values are the most interpretable.**

**Inference: The graph and table shown in problem 1 depict Kaplan-Meier estimates of survival probability for the 393 subjects with serum LDL less than 130mg/dL, the 225 subjects with LDL between 130mg/dL and 159mg/dL, and the 107 subjects with serum LDL of at least 160mg/dL. The graph and table show a trend for higher survival probabilities in groups with higher serum LDL levels. However, there are several points of crossover between the curves and table values. From proportional hazards regression analysis of 725 subjects, 131 of whom were observed to die during the study, we estimate that the observed differences in instantaneous risk death across groups defined by serum LDL modeled quadratically is greater than what might reasonably be expected when serum LDL had no true effect (p = 0.0061). Using a two-sided p value the second order term in serum LDL is not statistically significant at a 0.05 level of significance (p = 0.089) which suggests that we cannot with high confidence suggest that there is an association between instantaneous risk of death and groups defined by the square of serum LDL. Based on the overall significance of the model and the insignificant second order term we estimate that the association between instantaneous risk of death and serum LDL is linear, however, we cannot be certain about this conclusion.**

* 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 cldlsqr

predict fithrC

**The fitted hazard ratio variable was created using the predict function and the stata code listed above.**

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.



**The fitted values show a high level of variation at the endpoints but show a high level of agreement in the middle portion of the plot. The differences in the overall shapes of the fitted hazard ratio curves appears to reflect the different models that were used in problems 1 through 3. The fitted hazard ratios for serum LDL as a continuous variable seem to follow a linear trend. The fitted values for the log of serum LDL looks somewhat curvilinear and the fitted values for the serum LDL modeled quadratically appear to follow a u-shaped trend.**