**Homework #4**

1. A. Descriptive Statistics: Note that all descriptive statistics in this homework exclude the 10 patients who have missing LDL level observations. In this question, we are comparing the instantaneous hazard of death over the entire period of observation across groups defined by serum LDL modeled as a continuous variable. Since observation time is a censored variable, we created a Kaplan-Meier curve for survival estimates for all serum LDL levels.



In question 4, we will compare the instantaneous hazard of death over the entire period of observation by calculating the ratio of hazards between two groups differing in serum LDL level, where the hazard for the group with serum LDL not equal to 160 mg/dL is divided by the hazard for the group with serum LDL equal to 160 mg/dL. As this question dichotomizes serum LDL at 160 mg/dL, we also created a Kaplan-Meier curve which compares survival estimates between patients with serum LDL less than 160 mg/dL and those with serum LDL greater than or equal to 160 mg/dL.



We calculated the proportion of patients who survive to one, two, three, four, and five years among patients of all serum LDL levels, and also among patients with serum LDL less than 160 mg/dL as compared with patients with serum LDL greater than or equal to 160 mg/dL.

|  |  |
| --- | --- |
|  | Proportion Surviving (Kaplan-Meier) |
| Observation Time | Serum LDL<160 mg/dL (n=618) | Serum LDL>=160 mg/dL (n=107) | Any Serum LDL Level (n=725) |
| 1 year | 0.981 | 1.000 | 0.983 |
| 2 years | 0.952 | 0.981 | 0.956 |
| 3 years | 0.918 | 0.953 | 0.923 |
| 4 years | 0.887 | 0.907 | 0.890 |
| 5 years | 0.830 | 0.869 | 0.836 |

Note that in this study, less than 75% of patients of any serum LDL level were observed to die, and the same holds true among those patients with serum LDL less than 160 mg/dL and those with serum LDL greater than or equal to 160 mg/dL. For this reason, we calculated point estimates for the times by which 5%, 10%, 15%, and 20% of the sample was observed to die, for patients of any serum LDL level, and also for patients with serum LDL less than 160 mg/dL and those with serum LDL greater than or equal to 160 mg/dL.

|  |  |
| --- | --- |
|  | Time by Which Given Proportion of Sample Has Died (in months) |
| Proportion of the Sample Dead | Serum LDL<160 mg/dL (n=618) | Serum LDL>=160 mg/dL (n=107) | Any Serum LDL Level (n=725) |
| 5% | 24.0 | 37.2 | 26.1 |
| 10% | 43.5 | 49.5 | 44.0 |
| 15% | 56.1 | 64.7 | 56.8 |
| 20% | 65.3 | n/a | 66.4 |

We also calculated the restricted mean survival time for patients of any serum LDL level, and also for patients with serum LDL less than 160 mg/dL and those with serum LDL greater than or equal to 160 mg/dL. Since, in each of these groups, the last observation time is censored, these are underestimates of the true mean survival time.

|  |
| --- |
| Restricted Mean Survival Time (in months) |
| Serum LDL<160 mg/dL (n=618) | Serum LDL>=160 mg/dL (n=107) | Any Serum LDL Level (n=725) |
| 64.6 | 66.6 | 64.9 |

From each of the dichotomized descriptive statistics given above, it appears that patients with serum LDL greater than or equal to 160 mg/dL have a better survival experience (lower hazard of death) than patients with serum LDL less than 160 mg/dL.

Methods: We compared the instantaneous hazard of death over the entire period of observation across groups defined by serum LDL modeled as a continuous variable with a proportional hazards regression model with robust standard error estimates. Using Wald-based statistics, we found a 95% confidence interval for the ratio of hazards between two groups differing in serum LDL level by 10 mg/dL, where the hazard for the group with serum LDL 10 mg/dL higher is divided by the hazard for the group with the lower serum LDL. We also used Wald-based statistics to obtain a two-sided p-value to test the null hypothesis that the hazard of death does not vary across groups that differ in serum LDL level.Inferential Statistics and Results: From proportional hazards regression analysis with robust standard error estimates, we estimate that for each 10 mg/dL difference in serum LDL, the risk of death is 7.14% lower in the group with serum LDL 10 mg/dL higher as compared with the group with the lower serum LDL. This estimate is statistically significant at the 0.05 significance level with a two-sided Wald-based p-value of 0.009. Hence, we can with high confidence reject the null hypothesis that the hazard of death does not vary across groups that differ in serum LDL level in favor of the alternative hypothesis that the hazard of death does vary across groups that differ in serum LDL level. A Wald-based 95% confidence interval suggests that this observation would not be unusual if the risk of death were anywhere between 1.80% to 12.2% lower in the group with serum LDL 10 mg/dL higher as compared with the group with the lower serum LDL.

B. This has been computed and stored as fithrA. For graphs, see question 4.

2. A. Descriptive Statistics: In this question, we are comparing the instantaneous hazard of death over the entire period of observation across groups defined by serum LDL modeled as a logarithmically transformed continuous variable. For this reason, all descriptive statistics presented in part A of question 1 should be sufficient for this problem. Since the minimum serum LDL level is 11 mg/dL and since the natural logarithm is a monotonic function, when a descriptive statistic is computed for all serum LDL levels, this is the same as the descriptive statistic being computed for all logarithmically transformed serum LDL levels. For the same reason, when a descriptive statistic is computed for patients with serum LDL greater than or equal to 160 mg/dL and also for those with serum LDL less than 160 mg/dL, this is the same as that descriptive statistic being computed for patients with logarithmically transformed serum LDL greater than or equal to log(160)=5.08 and also for those with logarithmically transformed serum LDL less than 5.08.

Methods: We compared the instantaneous hazard of death over the entire period of observation across groups defined by serum LDL modeled as a logarithmically transformed continuous variable with a proportional hazards regression model with robust standard error estimates. Using Wald-based statistics, we found a 95% confidence interval for the ratio of hazards between two groups differing in serum LDL level by a factor of 2, where the hazard for the group with serum LDL twice as high is divided by the hazard for the group with the lower serum LDL. We also used Wald-based statistics to obtain a two-sided p-value to test the null hypothesis that the hazard of death does not vary across groups that differ in serum LDL level.Inferential Statistics and Results: From proportional hazards regression analysis with robust standard error estimates, we estimate that for each doubling in serum LDL, the risk of death is 43.6% lower in the group with serum LDL twice as high as compared with the group with the lower serum LDL. This estimate is statistically significant at the 0.05 significance level with a two-sided Wald-based p-value less than 0.001. Hence, we can with high confidence reject the null hypothesis that the hazard of death does not vary across groups that differ in serum LDL level in favor of the alternative hypothesis that the hazard of death does vary across groups that differ in serum LDL level. A Wald-based 95% confidence interval suggests that this observation would not be unusual if the risk of death were anywhere between 26.2% to 56.9% lower in the group with serum LDL twice as high as compared with the group with the lower serum LDL.

B. This has been computed and stored as fithrB. For graphs, see question 4.

3. A. Descriptive Statistics: In this question, we are comparing the instantaneous hazard of death over the entire period of observation across groups defined by serum LDL modeled quadratically as a continuous variable. For this reason, all descriptive statistics presented in part A of question 1 should be sufficient for this problem.Methods: We compared the instantaneous hazard of death over the entire period of observation across groups defined by serum LDL quadratically modeled as a continuous variable with a proportional hazards regression model with robust standard error estimates. We found a Z-based p-value to test if the association between serum LDL level and instantaneous hazard of death is nonlinear in a way that a quadratic polynomial could detect. We also used Wald-based statistics to obtain a two-sided Chi-Squared p-value to test the null hypothesis that the hazard of death does not vary across groups that differ in serum LDL level.Inferential Statistics and Results: We found a Z-based p-value of 0.0550 in testing whether the association between serum LDL level and instantaneous hazard of death is nonlinear in a way that a quadratic polynomial could detect, so we cannot with high confidence reject the null hypothesis that the association between serum LDL level and instantaneous hazard of death is linear or is nonlinear in a way that a quadratic polynomial could not detect. However, at the 0.05 significance level, a statistically significant two-sided Chi-Squared p-value of 0.0005 for the association between serum LDL level and instantaneous hazard of death suggests that we can with high confidence reject the null hypothesis that the hazard of death does not vary across groups that differ in serum LDL level in favor of the alternative hypothesis that the hazard of death does vary across groups that differ in serum LDL level.

B. This has been computed and stored as fithrC. For graphs, see question 4.

4.





The above graphs represent hazards for patients with any serum LDL level relative to the hazard for the group with serum LDL of 160 mg/dL under the various models from questions 1 through 3 on this homework. The model from question 1 has the relative hazard function fithrA, the model from question 2 has the relative hazard function fithrB, and the model from question 3 has the relative hazard function fithrC. The first graph displays these relative hazard functions on their full range of serum LDL levels, while the second graph displays the relative hazard functions only for serum LDL levels between 50 mg/dL and 200 mg/dL to show interesting intersections of these functions. For the models from each of questions 1 through 3, we can see that the relative hazard is greater than 1 for all serum LDL levels less than 160 mg/dL, and is equal to 1 for serum LDL equal to 160 mg/dL. For the models from questions 1 and 2, the relative hazard is less than 1 for all serum LDL levels greater than 160 mg/dL. For the model from question 3, however, the relative hazard is less than 1 for all serum LDL levels between 160 mg/dL and 184 mg/dL, is equal to 1 for serum LDL equal to 184 mg/dL, and is greater than 1 for all serum LDL levels greater than 160 mg/dL. For all serum LDL levels less than 18.9 mg/dL, the model from question 2 has the highest relative hazard, the model from question 3 has the second highest relative hazard, and the model from question 1 has the lowest relative hazard. For all serum LDL levels between 18.9 mg/dL and 74.2 mg/dL, the model from question 3 has the highest relative hazard, the model from question 2 has the second highest relative hazard, and the model from question 1 has the lowest relative hazard. For all serum LDL levels between 74.2 mg/dL and 86.1 mg/dL, the model from question 3 has the highest relative hazard, the model from question 1 has the second highest relative hazard, and the model from question 2 has the lowest relative hazard. For all serum LDL levels between 86.1 mg/dL and 97.2 mg/dL, the model from question 1 has the highest relative hazard, the model from question 3 has the second highest relative hazard, and the model from question 2 has the lowest relative hazard. For all serum LDL levels between 97.2 mg/dL and 160 mg/dL, the model from question 1 has the highest relative hazard, the model from question 2 has the second highest relative hazard, and the model from question 3 has the lowest relative hazard. For all serum LDL levels greater than 160 mg/dL, the model from question 3 has the highest relative hazard, the model from question 2 has the second highest relative hazard, and the model from question 1 has the lowest relative hazard.