**Homework #5**

1. A. Methods for descriptive statistics: Descriptive statistics for the distribution of diabetes diagnosis across race consist of a frequency table for these two variables. Note that there are no missing observations for either the race variable or the diabetes diagnosis variable.

Descriptive statistics results:

|  |  |
| --- | --- |
|  | Race |
| White (n=572) | Black (n=104) | Asian (n=47) | Other (n=12) |
| Diabetes Diagnosis | Diabetes | 56 | 18 | 3 | 2 |
| No Diabetes | 516 | 86 | 44 | 10 |

Methods for inferential statistics: The odds of subjects having diabetes were compared via odds ratios between subjects grouped by a dummy variable indicating the race of the subject, from among White (the reference group), Black, Asian, and Other, using a logistic regression model with robust standard error estimates. This is a saturated regression model, because we modeled the four race groups with three predictors and an intercept. Statistical inference was based on the Wald statistic computed from the regression predictor parameters and their standard errors, with two-sided p values and 95% confidence intervals computed for the intercept and each of the predictors using the approximate normal distribution for logistic regression parameter estimates.Inferential statistics results: For the 572 White subjects with available diabetes status, the estimated odds of having diabetes were 0.1085. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar White subjects were anywhere between 0.0824 and 0.1430. A two-sided p-value less than 0.0005 suggests that we can with 95% confidence reject the null hypothesis that the odds of a White subject having diabetes are 1 in favor of the alternative hypothesis that the odds of a White subject having diabetes are different from 1.For the 104 Black subjects with available diabetes status, the estimated odds of having diabetes were 0.2093. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar Black subjects were anywhere between 0.1174 and 0.3732. A two-sided p-value of 0.026 suggests that we can with 95% confidence reject the null hypothesis that the odds of a Black subject having diabetes are the same as the odds of a White subject having diabetes in favor of the alternative hypothesis that the odds of a Black subject having diabetes are different from the odds of a White subject having diabetes.For the 47 Asian subjects with available diabetes status, the estimated odds of having diabetes were 0.0682. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar Asian subjects were anywhere between 0.0205 and 0.2269. A two-sided p-value of 0.449 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Asian subject having diabetes are the same as the odds of a White subject having diabetes.For the 12 Other subjects with available diabetes status, the estimated odds of having diabetes were 0.2000. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar Other subjects were anywhere between 0.0427 and 0.9367. A two-sided p-value of 0.438 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Other subject having diabetes are the same as the odds of a White subject having diabetes.

B. The intercept in this model, once exponentiated, can be interpreted as the estimated odds of having diabetes among the White subjects. The predictor for the Black subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the Black subjects and the estimated odds of having diabetes among the White subjects, where the latter is divided by the former. The predictor for the Asian subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the Asian subjects and the estimated odds of having diabetes among the White subjects, where the latter is divided by the former. The predictor for the Other subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the Other subjects and the estimated odds of having diabetes among the White subjects, where the latter is divided by the former.

C. A two-sided p-value less than 0.0005 suggests that we can with 95% confidence reject the null hypothesis that the odds of a White subject having diabetes are 1 in favor of the alternative hypothesis that the odds of a White subject having diabetes are different from 1. A two-sided p-value of 0.026 suggests that we can with 95% confidence reject the null hypothesis that the odds of a Black subject having diabetes are the same as the odds of a White subject having diabetes in favor of the alternative hypothesis that the odds of a Black subject having diabetes are different from the odds of a White subject having diabetes. A two-sided p-value of 0.449 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Asian subject having diabetes are the same as the odds of a White subject having diabetes. A two-sided p-value of 0.438 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Other subject having diabetes are the same as the odds of a White subject having diabetes.

D. For descriptive statistics, see part A.Methods for inferential statistics: The odds of subjects having diabetes were compared via odds ratios between subjects grouped by a dummy variable indicating the race of the subject, from among Black (the reference group), White, Asian, and Other, using a logistic regression model with robust standard error estimates. This is a saturated regression model, because we modeled the four race groups with three predictors and an intercept. Statistical inference was based on the Wald statistic computed from the regression predictor parameters and their standard errors, with two-sided p values and 95% confidence intervals computed for the intercept and each of the predictors using the approximate normal distribution for logistic regression parameter estimates.Inferential statistics results: For the 104 Black subjects with available diabetes status, the estimated odds of having diabetes were 0.2093. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar Black subjects were anywhere between 0.1259 and 0.3480. A two-sided p-value less than 0.0005 suggests that we can with 95% confidence reject the null hypothesis that the odds of a Black subject having diabetes are 1 in favor of the alternative hypothesis that the odds of a Black subject having diabetes are different from 1.For the 572 White subjects with available diabetes status, the estimated odds of having diabetes were 0.1085. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar White subjects were anywhere between 0.0609 and 0.1935. A two-sided p-value of 0.026 suggests that we can with 95% confidence reject the null hypothesis that the odds of a White subject having diabetes are the same as the odds of a Black subject having diabetes in favor of the alternative hypothesis that the odds of a White subject having diabetes are different from the odds of a Black subject having diabetes.For the 47 Asian subjects with available diabetes status, the estimated odds of having diabetes were 0.0682. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar Asian subjects were anywhere between 0.0190 and 0.2442. A two-sided p-value of 0.085 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Asian subject having diabetes are the same as the odds of a Black subject having diabetes.For the 12 Other subjects with available diabetes status, the estimated odds of having diabetes were 0.2000. A 95% confidence interval suggests that these odds would not be unusual if the true odds of having diabetes among similar Other subjects were anywhere between 0.0403 and 0.9926. A two-sided p-value of 0.956 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Other subject having diabetes are the same as the odds of a Black subject having diabetes.

E. The intercept in this model, once exponentiated, can be interpreted as the estimated odds of having diabetes among the Black subjects. The predictor for the White subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the White subjects and the estimated odds of having diabetes among the Black subjects, where the latter is divided by the former. The predictor for the Asian subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the Asian subjects and the estimated odds of having diabetes among the Black subjects, where the latter is divided by the former. The predictor for the Other subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the Other subjects and the estimated odds of having diabetes among the Black subjects, where the latter is divided by the former.

F. A two-sided p-value less than 0.0005 suggests that we can with 95% confidence reject the null hypothesis that the odds of a Black subject having diabetes are 1 in favor of the alternative hypothesis that the odds of a Black subject having diabetes are different from 1. A two-sided p-value of 0.026 suggests that we can with 95% confidence reject the null hypothesis that the odds of a White subject having diabetes are the same as the odds of a Black subject having diabetes in favor of the alternative hypothesis that the odds of a White subject having diabetes are different from the odds of a Black subject having diabetes. A two-sided p-value of 0.085 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Asian subject having diabetes are the same as the odds of a Black subject having diabetes. A two-sided p-value of 0.956 suggests that we cannot with 95% confidence reject the null hypothesis that the odds of an Other subject having diabetes are the same as the odds of a Black subject having diabetes.

G. Note firstly that the p-values in part C lead to different conclusions than those in part F, since there are different reference groups in each model. In part C, the p-value for the White group parameter (the intercept) led to the conclusion that the odds of having diabetes in this group are different from 1, the p-value for the Black group parameter led to the conclusion that the odds of having diabetes are different between the White group and the Black group, and the p-values for the Asian group and the Other group each led to the conclusion that we cannot reject the null hypothesis that the odds of having diabetes in the White group are equivalent to the odds of having diabetes in the Asian group or the Other group, respectively. In part F, the p-value for the Black group parameter (the intercept) led to the conclusion that the odds of having diabetes in this group are different from 1, the p-value for the White group parameter led to the conclusion that the odds of having diabetes are different between the Black group and the White group, and the p-values for the Asian group and the Other group each led to the conclusion that we cannot reject the null hypothesis that the odds of having diabetes in the Black group are equivalent to the odds of having diabetes in the Asian group or the Other group, respectively. These are different, but related, conclusions, but in both C and F we obtain significant p-values for the White and Black groups and insignificant p-values for the Asian and Other groups. However, even this is misleading, since the p-values differ between parts C and F, and could conceivably differ to the point that a significant p-value from part C would have a corresponding insignificant p-value in part F, or vice versa. This means that if we used individual regression parameters from a dummy variable regression to decide whether to include or exclude those variables in a regression model, our decisions on which variables should be included or excluded in the model might vary depending on which variable is considered to correspond to the reference group, which is dangerous since our decisions on which variables are significant should not depend on which we choose as a reference variable. Especially note that in parts C and F, the variable which is chosen as the reference group has the lowest p-value, despite the fact that the reference variable has changed.

2. A. Methods for descriptive statistics: Descriptive statistics for the censoring distribution included the minimum and maximum observed censoring times and the Kaplan-Meier estimates of the 10th, 50th (median), and 90th percentiles, as well as the mean time of follow-up calculated as the area under the Kaplan-Meier estimate of the censoring distribution’s survivor curve.Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, standard deviation, and the 25th, 50th (median), and 75th percentiles for the cases with available data. For the purposes of descriptive statistics of the survival probabilities by serum LDL level, serum LDL was categorized according to the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Within these categories, Kaplan-Meier estimates of survival were calculated and graphed, and estimates of the 2 and 5 year survival probabilities, as well as the 10th and 20th percentiles of the survival distribution and the restricted mean survival during a period of observation that all LDL strata still had some subjects at risk (5.75 years).Descriptive statistics results: The study consisted of 735 subjects who were followed for death from any cause for a Kaplan-Meier estimated average of 5.33 years (median 5.66 years, range 5.00 to 5.91 years), during which time 133 deaths were observed. Serum LDL measurements at the time of study enrollment were not available on 10 subjects, two of whom were observed to die after 0.189 and 0.657 years of observation, with the remaining subjects still alive after 5.05 to 5.91 years of observation. In the 725 subjects with available serum LDL measurements at enrollment, the mean LDL was 126 mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL).Table 1 presents estimates of the survival distribution within strata defined by serum LDL and in the combined sample from the 725 subjects with available LDL measurements. The greatest difference in survival distributions is apparent when comparing those individuals having the lowest serum LDL levels (less than 70 mg/dL) at times after 2 years of follow-up. The 5 year survival probability is lowest in that group (59.1%) and is observed highest in the subjects having serum LDL between 160 and 189 mg/dL inclusive (88.0%). On average, the subjects in the lowest LDL stratum were estimated to average 4.91 years of life during the first 5.75 years following study enrollment, while the other strata averaged from 5.23 to 5.45 years. Figure 1 presents the Kaplan-Meier survival probability estimates graphically, where it is again the lowest LDL group that shows the most markedly different survival distribution.Table 1: Kaplan-Meier based estimates of distribution of time from study enrollment to death from any cause for subjects having serum LDL measurements at baseline.

|  |  |  |
| --- | --- | --- |
|  | Serum LDL at Study Enrollment |  |
|  | 11-69 mg/dL | 70-99 mg/dL | 100-129 mg/dL | 130-159 mg/dL | 160-189 mg/dL | 190-247 mg/dL | All Subjects (with LDL available) |
| N Subjects | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| N Deaths | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| 2 Year Survival Probability | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 96.7% |
| 5 Year Survival Probability | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 86.0% |
| 10th Pctile of Survival (in years) | 3.46 | 3.80 | 3.41 | 4.30 | 4.53 | 4.13 | 3.66 |
| 20th Pctile of Survival (in years) | 3.55 | 5.44 | 5.36 | NA | NA | NA | 5.54 |
| 5.75 Year Restricted Mean of Survival (in years) | 4.91 | 5.24 | 5.23 | 5.35 | 5.45 | 5.32 | 5.29 |

Footnote 1: Based on Kaplan-Meier estimates computed within strata defined by LDL and overall. NA indicates that the corresponding percentile is not estimable with the available data.Footnote 2: Average number of years alive during the first 5.75 years following study enrollment, as computed by the area under Kaplam-Meier survival curves computed within strata defined by LDL and overallFootnote 3: Ten of the 735 subjects in the study population were missing baseline serum LDL measurements. Two of those subjects were observed to die after 0.189 y and 0.657 years of observation. The remaining 8 subjects with missing LDL data were still alive at the end of their observation period 5.03 to 5.91 years after study enrollment.



Figure 1: Kaplan-Meier based estimates of distribution of time from study enrollment to death from any cause for 725 subjects having serum LDL measurements at baseline.

Methods for inferential statistics: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as a categorical dummy variable with intervals from 0-69 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Quantification of association between all cause mortality and LDL was summarized by the relative hazard ratios computed from the regression model, where the hazard of instantaneous death for the group of subjects with any serum LDL level greater than or equal to 70 mg/dL is divided by the hazard of instantaneous death for the group of subjects with any serum LDL level below 69 mg/dL. Confidence intervals and two-sided p values for estimated relative hazard ratios were computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.Inferential statistics results: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. Among the subjects with serum LDL level from 70-99 mg/dL, the estimated relative hazard ratio as compared to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL is 0.3980. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects with similar serum LDL 70-99 mg/dL as compared to subjects with serum LDL 0-69 mg/dL was anywhere between 0.2026017 and 0.7820043. A two-sided p-value of 0.008 suggests that we can with 95% confidence reject the null hypothesis that the hazard of instantaneous death among similar subjects with serum LDL 70-99 mg/dL is the same hazard as that for instantaneous death in the group with serum LDL level from 0-69 mg/dL in favor of the alternative hypotheses that these two hazards are not always equivalent.Among the subjects with serum LDL level from 100-129 mg/dL, the estimated relative hazard ratio as compared to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL is 0.3926. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects with similar serum LDL 100-129 mg/dL as compared to subjects with serum LDL 0-69 mg/dL was anywhere between 0.2071162 and 0.7441697. A two-sided p-value of 0.004 suggests that we can with 95% confidence reject the null hypothesis that the hazard of instantaneous death among similar subjects with serum LDL 100-129 mg/dL is the same hazard as that for instantaneous death in the group with serum LDL level from 0-69 mg/dL in favor of the alternative hypotheses that these two hazards are not always equivalent.Among the subjects with serum LDL level from 130-159 mg/dL, the estimated relative hazard ratio as compared to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL is 0.2939145. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects with similar serum LDL 130-159 mg/dL as compared to subjects with serum LDL 0-69 mg/dL was anywhere between 0.1521 and 0.5678. A two-sided p-value less than 0.0005 suggests that we can with 95% confidence reject the null hypothesis that the hazard of instantaneous death among similar subjects with serum LDL 130-159 mg/dL is the same hazard as that for instantaneous death in the group with serum LDL level from 0-69 mg/dL in favor of the alternative hypotheses that these two hazards are not always equivalent.Among the subjects with serum LDL level from 160-189 mg/dL, the estimated relative hazard ratio as compared to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL is 0.2565. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects with similar serum LDL 160-189 mg/dL as compared to subjects with serum LDL 0-69 mg/dL was anywhere between 0.113457 and 0.5799484. A two-sided p-value of 0.001 suggests that we can with 95% confidence reject the null hypothesis that the hazard of instantaneous death among similar subjects with serum LDL 160-189 mg/dL is the same hazard as that for instantaneous death in the group with serum LDL level from 0-69 mg/dL in favor of the alternative hypotheses that these two hazards are not always equivalent.Among the subjects with serum LDL level from 190-250 mg/dL, the estimated relative hazard ratio as compared to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL is 0.316718. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects with similar serum LDL 190-250 mg/dL as compared to subjects with serum LDL 0-69 mg/dL was anywhere between 0.1014077 and 0.9891787. A two-sided p-value of 0.048 suggests that we can with 95% confidence reject the null hypothesis that the hazard of instantaneous death among similar subjects with serum LDL 190-250 mg/dL is the same hazard as that for instantaneous death in the group with serum LDL level from 0-69 mg/dL in favor of the alternative hypotheses that these two hazards are not always equivalent.

B. The parameter for serum LDL at 70 mg/dL can be interpreted as the estimated relative hazard ratio comparing the hazard of instantaneous death among subjects with serum LDL level from 70-99 mg/dL to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL. The parameter for serum LDL at 100 mg/dL can be interpreted as the estimated relative hazard ratio comparing the hazard of instantaneous death among subjects with serum LDL level from 100-129 mg/dL to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL. The parameter for serum LDL at 130 mg/dL can be interpreted as the estimated relative hazard ratio comparing the hazard of instantaneous death among subjects with serum LDL level from 130-159 mg/dL to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL. The parameter for serum LDL at 160 mg/dL can be interpreted as the estimated relative hazard ratio comparing the hazard of instantaneous death among subjects with serum LDL level from 160-189 mg/dL to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL. The parameter for serum LDL at 190 mg/dL can be interpreted as the estimated relative hazard ratio comparing the hazard of instantaneous death among subjects with serum LDL level from 190-250 mg/dL to the hazard of instantaneous death among subjects with serum LDL level from 0-69 mg/dL.The intercept in this model, once exponentiated, can be interpreted as the estimated odds of having diabetes among the Black subjects. The predictor for the White subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the White subjects and the estimated odds of having diabetes among the Black subjects, where the latter is divided by the former. The predictor for the Asian subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the Asian subjects and the estimated odds of having diabetes among the Black subjects, where the latter is divided by the former. The predictor for the Other subject group, once exponentiated, can be interpreted as the ratio of the estimated odds of having diabetes among the Other subjects and the estimated odds of having diabetes among the Black subjects, where the latter is divided by the former.

C. To assess whether the regression model used in this problem provides a “better fit” than the model that uses only a continuous linear term for LDL, we first perform Cox regression on the model including both the continuous linear term for LDL as well each of the nonlinear predictors from the dummy variable. We next perform the Chi-Squared test on the nonlinear predictors and find a two-sided p-value of 0.3988, so we at 95% confidence, we cannot reject to the null hypothesis that the restricted model provides the best fit to this data.

D. This has been done, see Question 4.

3. A. For descriptive statistics, see question 2.Methods for inferential statistics: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as a linear spline variable with knots at 0 mg/dL, 70 mg/dL, 100 mg/dL, 130 mg/dL, 160 mg/dL, 190 mg/dL, and the maximum LDL level with non-missing observation time. Quantification of association between all cause mortality and LDL level was summarized by the relative hazard ratios computed from the regression model, where the hazard of instantaneous death for the group of subjects with any serum LDL level greater than or equal to 70 mg/dL is divided by the hazard of instantaneous death for the group of subjects with any serum LDL level below 69 mg/dL. Confidence intervals and two-sided p values for estimated relative hazard ratios were computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.Inferential statistics results: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. Among the subjects with serum LDL level equal to the scientifically irrelevant 0 mg/dL, the estimated hazard of instantaneous death is 0.9781. However, this number is more relevant for other reasons, since the estimated hazard of instantaneous death in any group with LDL between 0-70 mg/dL is 0.9871 the hazard of instantaneous death for a group with 1 mg/dL lower LDL. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects differing by 1 mg/dL of LDL between serum LDL 0-70 mg/dL was anywhere between 0.9602 and 0.9963. A two-sided p-value of 0.019 suggests that we can with 95% confidence reject the null hypothesis that for subjects with serum LDL levels from 0-70 mg/dL, the hazard of instantaneous death among subjects with 1 mg/dL higher serum LDL is the same as that for instantaneous death in the group with the lower serum LDL level. We reject this null hypothesis in favor of the alternative hypotheses that these two hazards are not always equivalent.The estimated hazard of instantaneous death in any group with LDL between 71-100 mg/dL is 0.9797 of the hazard of instantaneous death for a group with 1 mg/dL lower LDL. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects differing by 1 mg/dL of LDL between serum LDL 71-100 mg/dL was anywhere between 0.9535 and 1.007. A two-sided p-value of 0.139 suggests that we cannot with 95% confidence reject the null hypothesis that for subjects with serum LDL levels from 71-100 mg/dL, the hazard of instantaneous death among subjects with 1 mg/dL higher serum LDL is the same as that for instantaneous death in the group with the lower serum LDL level.The estimated hazard of instantaneous death in any group with LDL between 101-130 mg/dL is 0.9977 of the hazard of instantaneous death for a group with 1 mg/dL lower LDL. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects differing by 1 mg/dL of LDL between serum LDL 101-130 mg/dL was anywhere between 0.9764 and 1.019. A two-sided p-value of 0.835 suggests that we cannot with 95% confidence reject the null hypothesis that for subjects with serum LDL levels from 101-130 mg/dL, the hazard of instantaneous death among subjects with 1 mg/dL higher serum LDL is the same as that for instantaneous death in the group with the lower serum LDL level.The estimated hazard of instantaneous death in any group with LDL between 131-160 mg/dL is 1.004 of the hazard of instantaneous death for a group with 1 mg/dL lower LDL. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects differing by 1 mg/dL of LDL between serum LDL 131-160 mg/dL was anywhere between 0.9794 and 1.028. A two-sided p-value of 0.773 suggests that we cannot with 95% confidence reject the null hypothesis that for subjects with serum LDL levels from 131-160 mg/dL, the hazard of instantaneous death among subjects with 1 mg/dL higher serum LDL is the same as that for instantaneous death in the group with the lower serum LDL level.The estimated hazard of instantaneous death in any group with LDL between 161-190 mg/dL is 0.9709 of the hazard of instantaneous death for a group with 1 mg/dL lower LDL. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects differing by 1 mg/dL of LDL between serum LDL 161-190 mg/dL was anywhere between 0.9298 and 1.014. A two-sided p-value of 0.181 suggests that we cannot with 95% confidence reject the null hypothesis that for subjects with serum LDL levels from 161-190 mg/dL, the hazard of instantaneous death among subjects with 1 mg/dL higher serum LDL is the same as that for instantaneous death in the group with the lower serum LDL level.The estimated hazard of instantaneous death in any group with LDL greater than 190 mg/dL is 1.029 of the hazard of instantaneous death for a group with 1 mg/dL lower LDL. A 95% confidence interval suggests that this hazard ratio would not be unusual if the true hazard ratio comparing instantaneous death among subjects differing by 1 mg/dL of LDL greater than serum LDL of 190 mg/dL was anywhere between 0.9791 and 1.081. A two-sided p-value of 0.261 suggests that we cannot with 95% confidence reject the null hypothesis that for subjects with serum LDL levels greater than 190 mg/dL, the hazard of instantaneous death among subjects with 1 mg/dL higher serum LDL is the same as that for instantaneous death in the group with the lower serum LDL level.

B. The parameter for serum LDL at 0 mg/dL is essentially the intercept. It can be interpreted as the estimated hazard of instantaneous death among subjects with (an unrealistic) serum LDL level of 0 mg/dL. To determine the estimated hazard of instantaneous death among subjects with any serum LDL level “i” between 0 and 70 mg/dL, we can raise the parameter for serum LDL at 0 mg/dL to the serum LDL level “i”.The parameter for serum LDL at 70 mg/dL is can be interpreted as a sort of slope. To determine the estimated hazard of instantaneous death among subjects with any serum LDL level “i” between 71 and 100 mg/dL, we can raise the parameter for serum LDL at 0 mg/dL to the 70th power, raise the parameter for serum LDL at 70 mg/dL to the serum LDL level “i”, and then multiply these terms together.The parameter for serum LDL at 100 mg/dL is can be interpreted as a sort of slope. To determine the estimated hazard of instantaneous death among subjects with any serum LDL level “i” between 101 and 130 mg/dL, we can raise the parameter for serum LDL at 0 mg/dL to the 70th power, raise the parameter for serum LDL at 70 mg/dL to the 30th power, raise the parameter for serum LDL at 100 mg/L to the serum LDL level “i”, and then multiply these terms together.The parameter for serum LDL at 130 mg/dL is can be interpreted as a sort of slope. To determine the estimated hazard of instantaneous death among subjects with any serum LDL level “i” between 131 and 160 mg/dL, we can raise the parameter for serum LDL at 0 mg/dL to the 70th power, raise the parameter for serum LDL at 70 mg/dL to the 30th power, raise the parameter for serum LDL at 100 mg/L to the 30th power, raise the parameter for serum LDL at 130 mg/dL to the serum LDL level “i”, and then multiply these terms together.The parameter for serum LDL at 160 mg/dL is can be interpreted as a sort of slope. To determine the estimated hazard of instantaneous death among subjects with any serum LDL level “i” between 161 and 190 mg/dL, we can raise the parameter for serum LDL at 0 mg/dL to the 70th power, raise the parameter for serum LDL at 70 mg/dL to the 30th power, raise the parameter for serum LDL at 100 mg/L to the 30th power, raise the parameter for serum LDL at 130 mg/dL to the 30th power, raise the parameter for serum LDL at 160 mg/dL to the serum LDL level “i”, and then multiply these terms together.The parameter for serum LDL at 190 mg/dL is can be interpreted as a sort of slope. To determine the estimated hazard of instantaneous death among subjects with any serum LDL level “i” greater than or equal to 191, we can raise the parameter for serum LDL at 0 mg/dL to the 70th power, raise the parameter for serum LDL at 70 mg/dL to the 30th power, raise the parameter for serum LDL at 100 mg/L to the 30th power, raise the parameter for serum LDL at 130 mg/dL to the 30th power, raise the parameter for serum LDL at 160 mg/dL to the 30th power, raise the parameter for serum LDL at 190 mg/dL to the serum LDL level “i”, and then multiply these terms together.

C. To assess whether the regression model used in this problem provides a “better fit” than the model that uses only a continuous linear term for LDL, we first perform Cox regression on the model including each of the nonlinear predictors from the spline variable. We next perform the Chi-Squared test on the nonlinear predictors and find a two-sided p-value of 0.0788, so at 95% confidence, we cannot reject to the null hypothesis that the restricted model (which is a special case of linear spline-based regression) provides the best fit to this data.

D. This has been done, see Question 4.

4. A. The regression strategies used in Homeworks 4 and 5 used the full survival distribution by using observation times to describe all-cause mortality, while still taking censoring into account. However, the regression strategies in Homeworks 1-3 dichotomized the survival distribution by using only the binary variable of 5-year all-cause mortality to describe overall all-cause mortality. Since dichotomizing continuous variables involves loss of information about the survival distribution, and since the strategies in Homeworks 4 and 5 did not involve any loss of information about the survival distribution, the models from Homeworks 4 and 5 more accurately describe the survival distribution for patients.

B. Observe the labelled graph below, which plots all models from Homeworks 4 and 5 together. From this graph, we can see the same trends we noticed on Homework 4, where no single curve is always above the other curves. We can see that dummy variable curve fit does not follow a first-order linear trend, and on the far left, where there are most likely the smallest number of observations, the dummy variable curve is very different from the other curves. We also see that the linear spline model is very similar to the other 3 models from Homework 4 at the extreme ends, but it does have a couple of places in the center of the LDL level data where the linear spline curve is recognizably lower than the curves from Homework 4. On the whole, the graphs seem to be lining up relatively well considering how different the regression models were. Interestingly, in some regions, it appears as though the dummy variable function is the highest function, whereas the linear spline curve is the lowest. Overall, the linear spline model fits much better to the curves from Homework 4 than the dummy variable code.



C. A priori, of all the analyses we have considered for exploring an unadjusted association between all-cause mortality and serum LDL in an elderly population, I would prefer the regression analysis that involves all the information from the survival distribution. That is, I would prefer one of the methods from Homeworks 4 and 5 since none of those methods dichotomized survival, whereas all methods in Homeworks 1-3 involved dichotomizing the survival distribution by considering only 5-year mortality. I would also prefer a model that did not divide serum LDL level into various categories such as the methods in Homework 5, which might result in a loss of precision. With my own limited scientific knowledge of the relationship between serum LDL levels and survival, I have no strong reason to believe that there is a nonlinear trend between these two variables. Even if I did have knowledge of such a relationship, more simple linear or log linear regression models could accurately describe the average linear or log-linear trends present in the data. Hence, I would prefer not to use any of the methods from Homework 5, and I would also prefer not to use the method involving a quadratic LDL-based term from Homework 4. In comparing the two methods in Homework 4, I would prefer the method that modeled a continuous logarithmic transformation of serum LDL level, since this does not involve the loss of any information about the LDL level. With my limited understanding of biology, it seems that it is a doubling of proteins that can cause the most scientifically relevant outcome, as opposed to adding 2 units of the protein, so this is why I would choose the continuous logarithmically transformed serum LDL model over the continuous linear serum LDL model. Perhaps one of the greatest advantages of this model, however, is the fact that it is relatively straightforward to communicate and understand.