Problems 2 and 3 of the homework build on the analyses performed in homeworks #1 through #4. 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. Problem 1 of this homework uses the same dataset to explore associations between prevalence of diabetes and race in the population from which that sample was drawn.

1. Perform a statistical regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across.
   1. Fit a logistic regression model that uses whites as a reference group. Is this a saturated model? Provide a formal report (methods and inference) about the scientific question regarding an association between diabetes and race.

**This is a saturated model, we modeled 4 groups with 3 predictors plus an intercept (4 parameters) so the estimates will agree exactly with the sample odds.**

**Methods: The odds of subjects having diabetes were compared by fitting dummy variables across groups defined by race with categories of “White”, “Black”, “Asian”, and “Other”, using a logistic regression model. White was defined as the reference group. Statistical inference was based on the Wald statistic computed from the regression slope parameters and their standard errors, with two-sided p-values and 95% confidence intervals computed using the robust standard errors.**

**Inference: Of the 735 subjects, 572 were white, 104 were black, 47 were Asian and 12 were other. For the subjects who were white the odds of diabetes was 0.109, for the subjects who were black the odds of diabetes was 0.209, for the subjects who were Asian the odds of diabetes was 0.682, and for the subjects who identified as other the odds of diabetes was 0.200. Using the two-sided p-value for the entire model there is not a statically significant association (p = 0.0956) between odds of diabetes and racial group.**

* 1. Using the regression model fit in part (a), provide an interpretation for each of the regression parameters (including the intercept).

**There are four parameters in this model, 3 slopes and the intercept. The intercept tells the odds of diabetes for the reference group, whites in this model. Each slope tells the ratio of odds for diabetes comparing that racial group and the reference group. For this model the first slope tells the odds ratio for having diabetes comparing black and white subjects. The second slope tells the odds ratio for having diabetes comparing Asian and white subjects and the third slope tells the odds ratio for having diabetes comparing “other” and white subjects.**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Interpretation** | **Value** |
| **Intercept** | Odds of diabetes among white subjects | 0.1085 |
| **Coefficient 1** | Odds ratio for diabetes comparing black to white subjects | 1.9286 |
| **Coefficient 2** | Odds ratio for diabetes comparing Asian to white subjects | 0.6282 |
| **Coefficient 3** | Odds ratio for diabetes comparing “other” to white subjects | 1.8429 |

* 1. If we were to ignore issue related to multiple comparisons, what conclusions would you reach based on the p values reported in the regression output from part (a) using a 0.05 level of significance.

**Based on a 95% confidence interval, the observed odds ratio of 1.929 for comparison of the black subjects to the white subjects would not be judged unusual if the true odds ratio were anywhere between 1.082 to 3.439, with whites having lower odds of diabetes. Using a two-sided p-value this observation is statistically significant at a 0.05 level of significance (p = 0.026) and we reject the null hypothesis that there is no association between the odds of diabetes between white and black racial groups. Based on a 95% confidence interval, the observed odds ratio of 0.628 for comparison of the Asian subjects to the white subjects, with whites having lower odds of diabetes, would not be judged unusual if the true odds ratio were anywhere between 0.189 to 2.091. Using a two-sided p-value this observation is not statistically significant at a 0.05 level of significance (p = 0.449) and we fail to reject the null hypothesis that there is no association between the odds of diabetes between white and Asian racial groups. Based on a 95% confidence interval, the observed odds ratio of 1.843, with whites having lower odds for diabetes, for comparison of the “other” subjects to the white subjects would not be judged unusual if the true odds ratio were anywhere between 0.393 to 8.631. Using a two-sided p-value this observation is not statistically significant at a 0.05 level of significance (p = 0.438) and we fail to reject the null hypothesis that there is no association between the odds of diabetes between white and “other” racial groups. This would lead to the conclusion that there is an association between race and odds of diabetes because there is a significant association between black and white subjects and odds of diabetes.**

* 1. Now fit a logistic regression model that uses blacks as a reference group. How would your report of formal inference differ from that that you provided in part (a)? How does this regression model relate to that in part (a)?

**Methods: The odds of subjects having diabetes were compared by fitting dummy variables across groups defined by race with categories of “White”, “Black”, “Asian”, and “Other” using a logistic regression model. Black was defined as the reference group. Statistical inference was based on the Wald statistic computed from the regression slope parameters and their standard errors, with two-sided p-values and 95% confidence intervals computed using the robust standard errors.**

**Inference: Of the 735 subjects, 572 were white, 104 were black, 47 were Asian and 12 were other. For the subjects who were white the odds of diabetes was 0.109, for the subjects who were black the odds of diabetes was 0.209, for the subjects who were Asian the odds of diabetes was 0.682, and for the subjects who identified as other the odds of diabetes was 0.200. Using the two-sided p-value for the entire model there is not a statically significant association (p = 0.0956) between odds of diabetes and racial group.**

* 1. Using the regression model fit in part (d), provide an interpretation for each of the regression parameters (including the intercept.)

**There are four parameters in this model, 3 slopes and the intercept. The intercept tells the odds of diabetes for the reference group, blacks in this model. Each slope tells the ratio of odds for diabetes comparing that racial group and the reference group. For this model the first slope tells the odds ratio for having diabetes comparing black and white subjects. The second slope tells the odds ratio for having diabetes comparing Asian and black subjects and the third slope tells the odds ratio for having diabetes comparing “other” and black subjects.**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Interpretation** | **Value** |
| **Intercept** | Odds of diabetes among black subjects | 0.2093 |
| **Coefficient 1** | Odds ratio for diabetes comparing white to black subjects | 0.5185 |
| **Coefficient 2** | Odds ratio for diabetes comparing Asian to black subjects | 0.3256 |
| **Coefficient 3** | Odds ratio for diabetes comparing “other” to black subjects | 0.9556 |

* 1. If we were to ignore issue related to multiple comparisons, what conclusions would you reach based on the p values reported in the regression output from part (d) using a 0.05 level of significance.

**Based on a 95% confidence interval, the observed odds ratio of 0.519 for comparison of the black subjects to the white subjects would not be judged unusual if the true odds ratio were anywhere between 0.291 to 0.925 with white subjects having lower odds of diabetes. Using a two-sided p-value this observation is statistically significant at a 0.05 level of significance (p = 0.026) and we reject the null hypothesis that there is no association between the odds of diabetes between white and black racial groups. Based on a 95% confidence interval, the observed odds ratio of 0.326 for comparison of the Asian subjects to the black subjects, with Asians having lower odds of diabetes, would not be judged unusual if the true odds ratio were anywhere between 0.091 to 1.167. Using a two-sided p-value this observation is not statistically significant at a 0.05 level of significance (p = 0.085) and we fail to reject the null hypothesis that there is no association between the odds of diabetes between black and Asian racial groups. Based on a 95% confidence interval, the observed odds ratio of 0.956 for comparison of the “other” subjects to the black subjects, with “other” subjects having lower odds of diabetes, would not be judged unusual if the true odds ratio were anywhere between 0.193 to 4.742. Using a two-sided p-value this observation is not statistically significant at a 0.05 level of significance (p = 0.956) and we fail to reject the null hypothesis that there is no association between the odds of diabetes between black and “other” racial groups. This would lead to the conclusion that there is an association between race and odds of diabetes because there is a significant association between black and white subjects and odds of diabetes.**

* 1. What do your results from parts (c) and (f) say about the dangers of using the p values for individual regression parameters from a dummy variable regression to decide whether to include or exclude those variables in a regression model (i.e., in a “stepwise model building” procedure)?

**Associations can be made significant based on which group is chosen as the intercept. Therefore decisions to include or exclude variables in a regression model based on the individual parameters will be dependent on the coding that is used in the dummy regression which is a dangerous way to make decisions.**

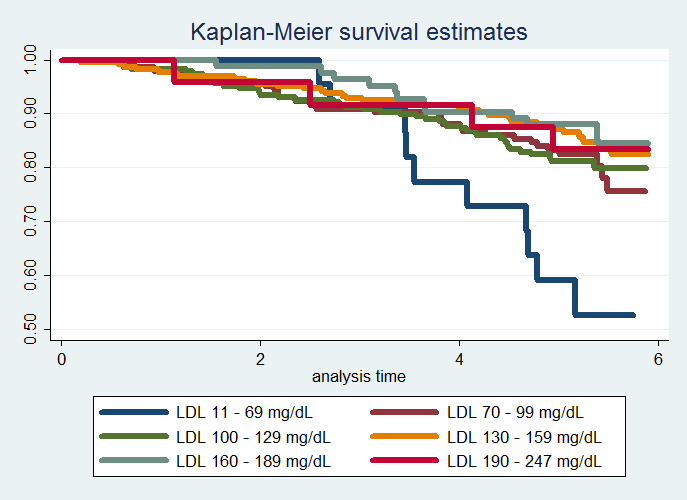
1. Perform a statistical regression analysis evaluating an association between all-cause mortality and serum by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL when fit as dummy variables using the categories suggested by the Mayo Clinic as reported on Homework #1. The Stata egen command can be used to categorize the LDL levels

egen ldlCTG = cut(ldl), at(0 70 100 130 160 190 250)

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

**Descriptive Statistics Methods: Descriptive statistics for the censoring distribution were the minimum and maximum observed censoring times and the mean time of follow-up calculated as the area under the Kaplan-Meier estimate of the censoring distribution’s survivor curve. The descriptive statistics for serum LDL levels were the number of missing cases the mean, standard deviation, and the min and max for cases with available data. To illustrate survival probability by serum LDL levels LDL was divided into categories using the Mayo Clinic guidelines: less than 70mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190mg/dL. Within these categories Kaplan-Meier estimates of the 2 and 5 year survival probabilities and the restricted means were calculated and the estimate of the survivor curve was graphed.**

**Descriptive Statistics: In this study there were 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 were taken at the time of study enrollment but were not available for 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. For the 725 subjects with available serum LDL measurements at time of enrollment, the mean LDL was 126mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL). The table below 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 seen when comparing individuals in the lowest serum LDL levels (less than 70 mg/dL) at times after 2 years of follow-up with other levels. The 5 year survival probability is lowest in the low LDL groups (59.1%) and is observed highest in the subjects having serum LDL between 160 and 189mg/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.**

****

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Serum LDL at Study Enrollment (mg/dL)** | | | | | | **All Subjects\*** |
|  | **11-69** | **70-99** | **100-129** | **130-159** | **160-189** | **190-247** |
| **Subjects** | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| **Deaths** | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| **2 year Survival Probability** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 95.6% |
| **5 year Survival Probability** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 83.6% |
| **5.75 year Restricted Mean Survival** | 4.91 | 5.24 | 5.23 | 5.35 | 5.45 | 5.32 | 5.29 |

**\*only subjects with LDL measurements at baseline were included**

**Inferential Statistics Methods: Distributions of time to death from any cause was compared across groups defined by serum LDL at enrollment using proportional hazards regression. Serum LDL was modeled using dummy variables fit across categories using the Mayo Clinic guidelines: less than 70mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190mg/dL. Quantification of association between all-cause mortality was summarized by the hazard ratios computed from the regression model, with confidence intervals and two-sided p-values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at baseline were not included in the analysis.**

**Inferential Results: Proportional hazards regression analysis was performed on 725 subjects having mean serum LDL of 126mg/dL (SD 33.6mg/dL; range 11-247mg/dL), 131 of whom were observed to die during the study with an average length of observation of 5.33 years. Using this model we estimate that there is a statistically significant (p = 0.0087) association between serum LDL and instantaneous risk of death. Because serum LDL was modeled categorically using dummy variables further inference about the instantaneous risk of death across categories cannot be carried out due to the type I error inflation associated with multiple comparisons.**

3/3 for descriptive statistics

3/3 for performing an appropriate analysis

1/4 for reporting the association appropriately

Wrong interpretation of coefficient (-2)

Wrong interpretation of CI (-1)

Total: 7

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.

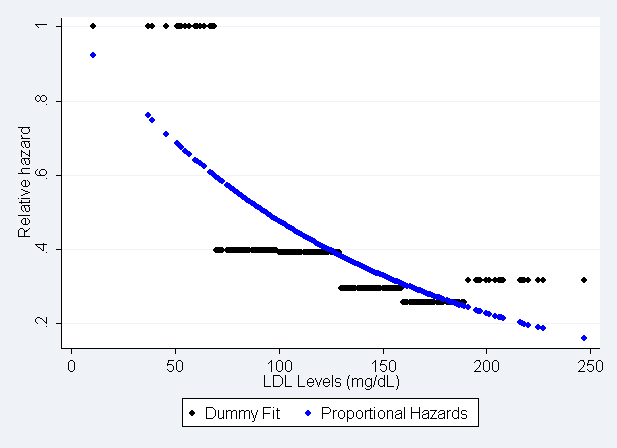
**There is no intercept in proportional hazards models, there is a baseline hazard function but it was not specified in this model. Each parameter gives the hazard ratio between groups differing by 1 unit in the predictor. For this model the groups are given by serum LDL categories divided using the Mayo Clinic guidelines: less than 70mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190mg/dL. The first parameter gives the hazard ratio for the group with serum LDL less than 70mg/dL and the group with serum LDL 70-99mg/dL. The second parameter gives the hazard ratio for the group with serum LDL between 70-99mg/dL and the group with serum LDL between 100-129mg/dL. The third parameter gives the hazard ratio for the group with serum LDL between 100-129mg/dL and the group with serum LDL between 130-159mg/dL. The fourth parameter gives the hazard ratio for the group with serum LDL between 130-159mg/dL and the group with serum LDL between 160-189mg/dL. The fifth parameter gives the hazard ratio for the group with serum LDL between 160-189mg/dL and the group with serum LDL of at least 190mg/dL.**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Interpretation** | **Value** |
| **Hazard ratio 1** | Hazard ratio for 5 year mortality between subjects with serum LDL <70mg/dL and those with LDL 70-99mg/dL | 0.3980 |
| **Hazard ratio 2** | Hazard ratio for 5 year mortality between subjects with serum LDL 70-99dL and those with LDL 100-129mg/dL | 0.3926 |
| **Hazard ratio 3** | Hazard ratio for 5 year mortality between subjects with serum LDL 100-129 and those with LDL 130-159mg/dL | 0.2939 |
| **Hazard ratio 4** | Hazard ratio for 5 year mortality between subjects with serum LDL 130-159mg/dL and those with LDL 160-189mg/dL | 0.2565 |
| **Hazard ratio 5** | Hazard ratio for 5 year mortality between subjects with serum LDL 160-189mg/dL and those with LDL ≥190mg/dL | 0.3167 |

Total: 5

* 1. What analysis would you perform to assess whether the regression model used in this problem provides a “better fit” than does a model that uses only a continuous linear term for LDL? What is the result of such an analysis?

**Methods: In order to assess whether the proportional hazards regression model using dummy variables for serum LDL as a categorical variable provides a “better fit” than the proportional hazards model using a continuous linear term for serum LDL a plot was generated comparing fitted hazard ratios generated using each model.**

****

**Results: Both models yield significant results for an association between serum LDL and instantaneous risk of death. However, as shown in the figure above the categorical dummy variable fit does not accurately predict the hazard ratios for values within the categories. In almost all cases the dummy fit over or underestimates the hazard ratio. As is shown in the figure above the continuous fit is much smoother and is a better reflection of the overall trend in the data.**

Did not mention including linear term (-1)

Did not mention about the test that regression coefficients for the dummy variables were 0 (-1)  
Did not mention what kind of test you use (-1)  
Wrong p-value (-1)  
Wrong conclusion (-1)

Total: 0

* 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). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.

**The variable was created; fitted variables will be shown in problem 4.**

1. Perform a statistical regression analysis evaluating an association between all-cause mortality and serum by comparing the instantaneous risk (hazard) of death over the entire period of observation across groups defined by serum LDL when fit as linear splines using the categories suggested by the Mayo Clinic as reported on Homework #1. The Stata mkspline command can be used to create the predictors that can be used in a regression

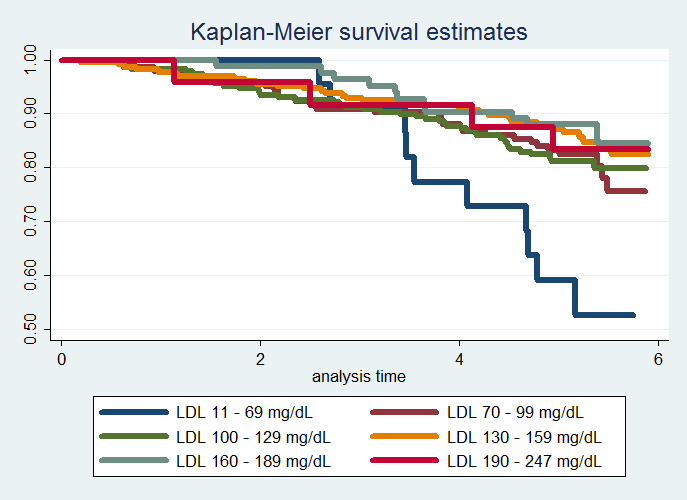
mkspline ldl0 70 ldl70 100 ldl100 130 ldl130 160 ldl160 190 ldl190 = ldl

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

(Descriptive statistics are the same as those generated for problem 2a)

**Descriptive Statistics Methods: Descriptive statistics for the censoring distribution were the minimum and maximum observed censoring times and the mean time of follow-up calculated as the area under the Kaplan-Meier estimate of the censoring distribution’s survivor curve. The descriptive statistics for serum LDL levels were the number of missing cases the mean, standard deviation, and the min and max for cases with available data. To illustrate survival probability by serum LDL levels LDL was divided into categories using the Mayo Clinic guidelines: less than 70mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190mg/dL. Within these categories Kaplan-Meier estimates of the 2 and 5 year survival probabilities and the restricted means were calculated and the estimate of the survivor curve was graphed.**

**Descriptive Statistics: In this study there were 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 were taken at the time of study enrollment but were not available for 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. For the 725 subjects with available serum LDL measurements at time of enrollment, the mean LDL was 126mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL). The table below 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 seen when comparing individuals in the lowest serum LDL levels (less than 70 mg/dL) at times after 2 years of follow-up with other levels. The 5 year survival probability is lowest in the low LDL groups (59.1%) and is observed highest in the subjects having serum LDL between 160 and 189mg/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.**

****

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Serum LDL at Study Enrollment (mg/dL)** | | | | | | **All Subjects\*** |
|  | **11-69** | **70-99** | **100-129** | **130-159** | **160-189** | **190-247** |
| **Subjects** | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| **Deaths** | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| **2 year Survival Probability** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 95.6% |
| **5 year Survival Probability** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 83.6% |
| **5.75 year Restricted Mean Survival** | 4.91 | 5.24 | 5.23 | 5.35 | 5.45 | 5.32 | 5.29 |

**\*only subjects with LDL measurements at baseline were included**

**Inferential Statistics Methods: Distributions of time to death from any cause was compared across groups defined by serum LDL at enrollment using proportional hazards regression. Serum LDL was modeled using a spline fit with knots at the cut points for serum LDL categories according to the Mayo Clinic guidelines: 70mg/dL, 100 mg/dL, 130 mg/dL, 160mg/dL, and 190mg/dL. Quantification of association between all-cause mortality was summarized by the hazard ratios computed from the regression model, with confidence intervals and two-sided p-values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at baseline were not included in the analysis.**

**Inferential Results: Proportional hazards regression analysis was performed on 725 subjects having mean serum LDL of 126mg/dL (SD 33.6mg/dL; range 11-247mg/dL), 131 of whom were observed to die during the study with an average length of observation of 5.33 years. Using this model we estimate that there is a statistically significant (p < 0.0001) association between serum LDL and instantaneous risk of death. Because serum LDL was modeled categorically using a spline fit further inference about the instantaneous risk of death across categories cannot be carried out due to the type I error inflation associated with multiple comparisons.**

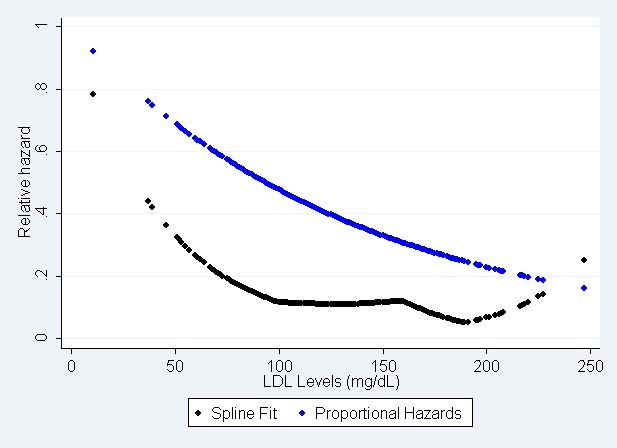
* 1. Provide an interpretation for each parameter in your regression model, including the intercept.

**There is no intercept in proportional hazards models, there is a baseline hazard function but it was not specified in this model. Each parameter gives the hazard ratio between two groups both between the same knots but differing by 1 unit in the predictor. For this model the groups are given by a spline fit of serum LDL with knots at the cut points for serum LDL categories according to the Mayo Clinic guidelines: 70mg/dL, 100 mg/dL, 130 mg/dL, 160mg/dL, and 190mg/dL. The first parameter gives the hazard ratio for each 1mg/dL increase in serum LDL for subjects with serum LDL levels between 0 and 69mg/dL at baseline. The second parameter gives the hazard ratio for each 1mg/dL increase in serum LDL for subjects with serum LDL levels between 70 and 99mg/dL at baseline. The third parameter gives the hazard ratio for each 1mg/dL increase in serum LDL for subjects with serum LDL levels between 100 and 129mg/dL at baseline. The fourth parameter gives the hazard ratio for each 1mg/dL increase in serum LDL for subjects with serum LDL levels between 130 and 159mg/dL at baseline. The fifth parameter gives the hazard ratio for each 1mg/dL increase in serum LDL for subjects with serum LDL levels between 160 and 189mg/dL at baseline. The sixth parameter gives the hazard ratio for each 1mg/dL increase in serum LDL for subjects with serum LDL levels of at least 190mg/dL at baseline.**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Interpretation** | **Value** |
| **Hazard ratio 1** | Hazard ratio for 5 year mortality for each 1mg/dL increase in serum LDL for subjects with serum LDL from 0-69mg/dL | 0.9781 |
| **Hazard ratio 2** | Hazard ratio for 5 year mortality for each 1mg/dL increase in serum LDL for subjects with serum LDL from 70-99mg/dL | 0.9797 |
| **Hazard ratio 3** | Hazard ratio for 5 year mortality for each 1mg/dL increase in serum LDL for subjects with serum LDL from 99-129mg/dL | 0.9977 |
| **Hazard ratio 4** | Hazard ratio for 5 year mortality for each 1mg/dL increase in serum LDL for subjects with serum LDL from 129-159mg/dL | 1.0036 |
| **Hazard ratio 5** | Hazard ratio for 5 year mortality for each 1mg/dL increase in serum LDL for subjects with serum LDL from 160-189mg/dL | 0.9709 |
| **Hazard ratio 6** | Hazard ratio for 5 year mortality for each 1mg/dL increase in serum LDL for subjects with serum LDL ≥190mg/dL | 1.0288 |

* 1. What analysis would you perform to assess whether the regression model used in this problem provides a “better fit” than does a model that uses only a continuous linear term for LDL? What is the result of such an analysis?

**Methods: In order to assess whether the proportional hazards regression model using a spline fit of serum LDL as a categorical variable provides a “better fit” than the proportional hazards model using a continuous linear term for serum LDL a plot was generated comparing fitted hazard ratios generated using each model.**

****

**Results: Both models yield significant results for an association between serum LDL and instantaneous risk of death. However, as shown in the figure above the spline fit does not accurately predict the hazard ratios for values within the categories. In almost all cases the spline fit underestimates the hazard ratio. As is shown in the figure above the continuous fit is much smoother and is a better reflection of the overall trend in the data.**

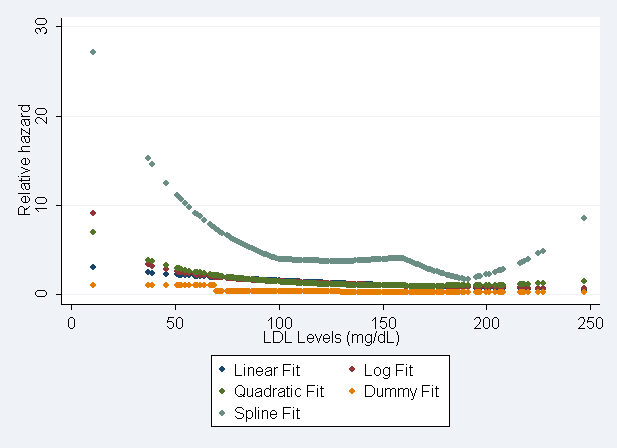
* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.

**The variable was created; fitted variables will be shown in problem 4.**

1. By answering the following questions, compare the relative advantages and disadvantages of the various statistical analysis strategies we have considered in Homeworks 1-4 and problems 2 and 3 in this homework.
   1. What advantages do the regression strategies used in Homeworks 4 and 5 provide over the approaches used in Homeworks 1-3?

**For homeworks 4 and 5 proportional hazard regression models are used. This allow us to analyze censored data so observation time, time to censoring or time to death, can be analyzed as a continuous variable using Kaplan-Meier methods which allows for more precision in the analysis. In homeworks 1-3 observation time was truncated at 5 years to avoid the issue of censoring. By only looking at whether subjects were alive or dead at 5 years we decreased the precision of our analysis to detect time trends in mortality across serum LDL groups.**

* 1. Comment on any similarities or differences of the fitted values from the three models fit in Homework 4 and the two models fit in problems 2 and 3 of this homework.



**The fitted values for the three models fit in homework 4 as well as those for the two models fit in this homework are shown in the above figure. In each case the models show a general downward trend with higher serum LDL levels having increased risk of instantaneous death. There is not a large difference in the linear, quadratic, log and dummy fits, especially in the middle range of the data. The spline fit shows significant deviation from the rest of the models. All models show large differences in the low range of serum LDL values, reflecting the sparseness of the data in this range. There is also poor agreement in the high range of the data, also reflecting the availability of few data points. The fits from the logarithmic and quadratic fits remain remarkably similar. If these similarities do reflect an accurate representation of the true association the logarithmic fit is still the preferable model as it has better interpretability. A larger sample would be needed to establish if there is a true u-shape trend in instantaneous risk of death and serum LDL with higher and lower values of LDL having increased risk.**

* 1. *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, which one would you prefer and why?

***A priori* I would prefer to use Kaplan-Meier methods and proportional hazards regression on the data as this would allow me to have greater precision in my analysis by leaving time to death/censoring as a continuous variable. I would also prefer to treat serum LDL as a continuous variable, again to maintain precision, and I would use the log transformed serum LDL values in the model. Because we are looking at a predominantly healthy elderly adult population we would not expect serum LDL to be over a huge range of values, however, if there is a multiplicative trend in the association between serum LDL and all-cause mortality then a log model would better capture that trend.**