**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #5**

February 3, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 10, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is* ***TOTALLY*** *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

***Unless explicitly told otherwise in the statement of the problem, in all problems requesting “statistical analyses” (either descriptive or inferential), you should present both***

* ***Methods: A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

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.

**Methods**. In order to assess the potential association between diabetes and race, we estimate the odds ratio of diabetes diagnosis by race, fitting a logistic regression model with whites as the reference group. All subjects are categorized as being white, black, Asian, or other for a total of 4 groups. The logistic regression we fit includes an intercept, estimating the odds for whites, and three slope coefficients that estimate the odds ratios for each of the non-reference racial groups. Since the model has equal numbers of parameters and groups, it is saturated. Standard errors for the model were calculated using the Huber-White sandwich estimator.

**Inference**. Of the 735 individuals in the data set, 79 had received diagnosis of diabetes. Of the subjects diagnosed with diabetes, 56 were white, 18 were black, 3 were Asian and 2 were other. The 656 individuals that were not diagnosed with diabetes consisted of 516 whites, 86 blacks, 44 Asians, and 10 other.

From the logistic regression model, we estimate that the odds of white individuals being diagnosed with diabetes are 0.1085. Based on the estimated odds, the risk of diabetes diagnosis given that the subject is white is 9.79% (0.1085/ (1+0.1085) = 0.0979). The observed odds of diabetes diagnosis amongst white individuals would be consistent with a true odds between 0.0824 to 0.1430 based on the 95% (robust) confidence interval.

From the logistic model, we obtained estimates of the odds ratios for each of blacks, Asians, and other races relative to a white population.

For blacks, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 92.86% higher relative to the odds in an otherwise similar population of white individuals. This estimate would be consistent with a true odds that is 8.15% higher to 243.9% higher relative to the odds in an otherwise similar population of white individuals based on a robust 95% confidence interval.

For Asians, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 37.18% lower relative to the odds in an otherwise similar population of white individuals. This estimate would be consistent with true odds that are 81.12% lower to 109.1% higher relative to the odds in an otherwise similar population of white individuals based on a robust 95% confidence interval.

For other races, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 84.29% higher relative to the odds in an otherwise similar population of white individuals. This estimate would be consistent with true odds that are 60.65% lower to 763.1% higher relative to the odds in an otherwise similar population of white individuals based on a robust 95% confidence interval.

The chi-square test that all of the parameters in the model are equal to zero had a p-value of 0.0956. Hence we do not have strong evidence of an association between race and diabetes.

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

**Answer**. Setting all dummy variables for race equal to zero, the intercept of the logistic regression model we fit above provides an estimate of the odds of a white individual to receive a diabetes diagnosis.

There are three slope parameters, one for each of blacks, Asians, and others. The slope parameters correspond to the estimated odds ratio for diabetes diagnosis of that race relative to whites.

For blacks, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 92.86% higher relative to the odds in an otherwise similar population of white individuals.

For Asians, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 37.18% lower relative to the odds in an otherwise similar population of white individuals.

For other races, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 84.29% higher relative to the odds in an otherwise similar population of white individuals.

* 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.

**Answer**. For blacks, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 92.86% higher relative to the odds in an otherwise similar population of white individuals. This estimate would be consistent with a true odds that is relatively 8.15% higher to 243.9% higher based on a robust 95% confidence interval. The test that the odds ratio is equal to 1 (a relative increase/decrease of 0%) has a p-value of 0.026, so we have statistically significant evidence at the 0.05 confidence level that the odds of blacks receiving a diabetes diagnosis are higher than the same odds in a white population.

For Asians, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 37.18% lower relative to the odds in an otherwise similar population of white individuals. This estimate would be consistent with true odds that are relatively 81.12% lower to 109.1% higher based on a robust 95% confidence interval. The test that the odds ratio is equal to 1 (a relative increase/decrease of 0%) has a p-value of 0.449, so we do not have statistically significant evidence at the 0.05 confidence level that the odds of Asians receiving a diabetes diagnosis differ from those same odds in a white population.

For other races, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 84.29% higher relative to the odds in an otherwise similar population of white individuals. This estimate would be consistent with true odds that are relatively 60.65% lower to 763.1% higher based on a robust 95% confidence interval. The test that the odds ratio is equal to 1 (a relative increase/decrease of 0%) has a p-value of 0.438, so we do not have statistically significant evidence at the 0.05 confidence level that the odds of non-white, non-black, non-Asians receiving a diabetes diagnosis differ from those same odds in a white population.

* 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**. In order to assess the potential association between diabetes and race, we estimate the odds ratio of diabetes diagnosis by race, fitting a logistic regression model with blacks as the reference group. All subjects are categorized as being white, black, Asian, or other for a total of 4 groups. The logistic regression we fit includes an intercept, estimating the odds for blacks, and three slope coefficients that estimate the odds ratios for each of the non-reference racial groups. Since the model has equal numbers of parameters and groups, it is saturated. Standard errors for the model were calculated using the Huber-White sandwich estimator.

**Inference**. Our inference based on this model is still that the chi-square test that all of the parameters in the model are equal to zero had a p-value of 0.0956. Hence we do not have strong evidence of an association between race and diabetes.

Like in the model from (a), this is a saturated model estimating the odds, so the regression intercept will be exactly the odds of blacks having a diabetes diagnosis and the slopes are the odds ratios of whites, Asians and other races having a diabetes diagnosis relative to blacks. This differs from the model in (a) because the reference group in (a) was whites instead of blacks.

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

**Answer**. Setting all dummy variables for race equal to zero, the intercept of the logistic regression model provides an estimate that the odds of a black individual to receive a diabetes diagnosis are 0.2093.

There are three slope parameters, one for each of whites, Asians, and others. The slope parameters correspond to the estimated odds ratio for diabetes diagnosis of that race relative to blacks.

For whites, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 48.15% lower relative to the odds in an otherwise similar population of black individuals.

For Asians, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 67.42% lower relative to the odds in an otherwise similar population of black individuals.

For other races, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 4.44% lower relative to the odds in an otherwise similar population of black individuals.

* 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.

**Answer.** For whites, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 48.15% lower relative to the odds in an otherwise similar population of black individuals. This estimate would be consistent with true odds that are relatively 70.92% lower to 7.54% lower based on a robust 95% confidence interval. The test that the odds ratio is equal to 1 (a relative increase/decrease of 0%) has a p-value of 0.026, so we have statistically significant evidence at the 0.05 confidence level that the odds of whites receiving a diabetes diagnosis are lower than those same odds in a black population.

For Asians, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 67.42% lower relative to the odds in an otherwise similar population of black individuals. This estimate would be consistent with true odds that are relatively 90.91% lower to 16.69% higher based on a robust 95% confidence interval. The test that the odds ratio is equal to 1 (a relative increase/decrease of 0%) has a p-value of 0.085, so we do not have statistically significant evidence at the 0.05 confidence level that the odds of Asians receiving a diabetes diagnosis differ from those same odds in a black population.

For other races, the slope from the logistic regression estimates that the odds of a diabetes diagnosis are 4.44% lower relative to the odds in an otherwise similar population of black individuals. This estimate would be consistent with true odds that are relatively 80.75% lower to 374.2% higher based on a robust 95% confidence interval. The test that the odds ratio is equal to 1 (a relative increase/decrease of 0%) has a p-value of 0.956, so we do not have statistically significant evidence at the 0.05 confidence level that the odds of non-white, non-black, non-Asians receiving a diabetes diagnosis differ from those same odds in a black population.

* 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)?

**Answer**. In part (c), comparisons between groups are made with the odds ratio relative to a white population. In part (f), comparison between groups are made with the odds ratio relative to a black population. While the p-value for the comparison of whites to blacks (or blacks to whites) does not change between these models, the other parameters are not estimating the same comparison anymore and thus have wildly different p-values (though their non-significance at the 0.05 level has not changed). This shows that depending on our reference group, we may come to different conclusions about the significance of observed associations when we fit future models with dummy variables.

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.

**Methods**. In order to assess the potential association between serum LDL and all-cause mortality, we estimate the hazards ratio across categories defined by the Mayo clinic’s recommendations for serum LDL (below 70 mg/dL, below 100 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, 190 mg/dL and above), fitting a Cox proportional hazards regression model with serum LDL below 70 mg/dL as the reference group. Standard errors for the model are calculated using the Huber-White sandwich estimator. The chi-square test with null hypothesis that all hazard ratios are equal to 1 will be used to determine whether we have significant evidence of association between serum LDL and all-cause mortality.

**Descriptives**. Of 735 subjects on the study, there were 10 with missing serum LDL measurements. There are 22 individuals with serum LDL of 11-69 mg/dL, 143 with serum LDL from 70-99 mg/dL, 228 with serum LDL from 100-129 mg/dL, 225 with serum LDL from 130-159 mg/dL, 83 with serum LDL from 160-189 mg/dL, and 24 individuals with serum LDL from 190-247 mg/dL. Due to the right-censoring present in the survival times of the subjects, we compare the Kaplan-Meier estimates of the survival curves across groups according to the aforementioned Mayo Clinic suggestions. Based on the plot, the group with LDL under 70 mg/dL had the worst survival over the full time period. Other ranges of LDL appeared to have more similar survival estimates over the study period.



**Inference**. From fitting a Cox proportional hazards regression model with serum LDL below 70 mg/dL as the reference group, we estimate hazards ratios of the other LDL categories relative to this group. For subjects with serum LDL of 70-99 mg/dL, the hazard is estimated to be 60.2% lower relative to a group with serum LDL less than 70 mg/dL.

For subjects with serum LDL of 100-129 mg/dL, the hazard is estimated to be 60.74% lower relative to a group with serum LDL less than 70 mg/dL. This would be consistent with a true hazard that is from 79.74% lower to 21.80% lower relative to a group with serum LDL less than 70 mg/dL based on a robust 95% confidence interval.

For subjects with serum LDL of 130-159 mg/dL, the hazard is estimated to be 70.61% lower relative to a group with serum LDL less than 70 mg/dL. This would be consistent with a true hazard that is from 79.29% lower to 25.58% lower relative to a group with serum LDL less than 70 mg/dL based on a robust 95% confidence interval.

For subjects with serum LDL of 160-189 mg/dL, the hazard is estimated to be 74.35% lower relative to a group with serum LDL less than 70 mg/dL. This would be consistent with a true hazard that is from 84.79% lower to 43.22 % lower relative to a group with serum LDL less than 70 mg/dL based on a robust 95% confidence interval.

For subjects with serum LDL of at least 190 mg/dL, the hazard is estimated to be 68.33% lower relative to a group with serum LDL less than 70 mg/dL. This would be consistent with a true hazard that is from 89.86% lower to 1.08% lower relative to a group with serum LDL less than 70 mg/dL based on a robust 95% confidence interval.

The chi-square test with the hypothesis that all of the true hazard ratios are equal to one has a p-value of 0.0087, so we have statistically significant evidence to reject this null hypothesis in favor of a hypothesis that there is at least one hazard ratio that is less than one.

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

**Answer**. No intercept is available for the Cox regression model, as we are not generally interested in the baseline hazard. We do fit slope parameters for the categories serum LDL of 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and 190+ mg/dL. Setting one of the dummy variables to one and all other covariates to zero provides an estimate of the hazard ratio of that category to the reference group, which has serum LDL below 70 mg/dL.

For subjects with serum LDL of 70-99 mg/dL, the hazard is estimated to be 60.2% lower relative to a group with serum LDL less than 70 mg/dL.

For subjects with serum LDL of 100-129 mg/dL, the hazard is estimated to be 60.74% lower relative to a group with serum LDL less than 70 mg/dL.

For subjects with serum LDL of 130-159 mg/dL, the hazard is estimated to be 70.61% lower relative to a group with serum LDL less than 70 mg/dL.

For subjects with serum LDL of 160-189 mg/dL, the hazard is estimated to be 74.35% lower relative to a group with serum LDL less than 70 mg/dL.

For subjects with serum LDL of at least 190 mg/dL, the hazard is estimated to be 68.33% lower relative to a group with serum LDL less than 70 mg/dL.

* 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**. To assess whether the Cox proportional hazards regression model in (a) provides a better fit than a model using only a continuous linear term for LDL, we fit a new Cox regression model that includes a continuous linear term for LDL as a predictor in addition to the dummy variables for the categories defined by the Mayo clinic’s recommendations for serum LDL. We then test the hypothesis that all the dummy variables are equal to zero with a partial chi-square test.

**Inference**. From Cox proportional hazards regression, the partial chi-square test with null hypothesis that the slope parameters for the Mayo clinic categories of serum LDL are all zero has a p-value of 0.3988. We do not have statistically significant evidence of a non-linear trend in the hazard ratio as predicted by serum LDL, thus a model using only a continuous linear term for LDL would provide a better fit. Despite this, we cannot rule out the possibility of a true non-linear association based on this analysis alone.

* 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.

**Answer**. See graph for 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.

**Methods**. In order to assess the potential association between serum LDL and all-cause mortality, we estimate the hazards ratio by Cox proportional hazards regression fit as linear splines with knots using the categories suggested by the Mayo Clinic for serum LDL (below 70 mg/dL, below 100 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, 190 mg/dL and above). Standard errors for the model are calculated using the Huber-White sandwich estimator. The chi-square test with null hypothesis that all hazard ratios are equal to 1 will be used to determine whether we have significant evidence of association between serum LDL and all-cause mortality.

**Descriptives**. See question 2 for discussion of these.

**Inference**. For two groups of individuals with serum LDL below 70 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.19% lower on average. This estimate would be consistent with a true hazard that is between 3.98% lower to 0.37% lower based on a robust 95% confidence interval.

For two groups of individuals with serum LDL between 70-99 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.03% lower on average. This estimate would be consistent with a true hazard that is between 4.65% lower to 0.67% higher based on a robust 95% confidence interval.

For two groups of individuals with serum LDL between 100-129 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 0.23% lower on average. This estimate would be consistent with a true hazard that is between 2.36% lower to 1.95% higher based on a robust 95% confidence interval.

For two groups of individuals with serum LDL between 130-159 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 0.36% higher on average. This estimate would be consistent with a true hazard that is between 2.06% lower to 2.84% higher based on a robust 95% confidence interval.

For two groups of individuals with serum LDL between 160-189 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.91% lower on average. This estimate would be consistent with a true hazard that is between 7.02% lower to 1.38% higher based on a robust 95% confidence interval.

For two groups of individuals with serum LDL of at least 190 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.88% higher on average. This estimate would be consistent with a true hazard that is between 2.09% lower to 8.10% higher based on a robust 95% confidence interval.

The chi-squared test with the hypothesis that all of the hazard ratios estimated are equal to 1 has a p-value of less than 0.0001 (chi-square statistic 31.77 with 6 degrees of freedom). Hence we have statistically significant evidence that there is an association between serum LDL and all-cause mortality.

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

**Answer**. Our model fit a slope parameter for each of the following regions of serum LDL: below 70 mg/dL, 70-100 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and at least 190 mg/dL. In each case, the slope is the hazard ratio between groups differing in serum LDL by 1 mg/dL within the same LDL interval.

For two groups of individuals with serum LDL below 70 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.19% lower on average.

For two groups of individuals with serum LDL between 70-99 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.03% lower on average.

For two groups of individuals with serum LDL between 100-129 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 0.23% lower on average.

For two groups of individuals with serum LDL between 130-159 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 0.36% higher on average.

For two groups of individuals with serum LDL between 160-189 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.91% lower on average.

For two groups of individuals with serum LDL of at least 190 mg/dL that differ in serum LDL by 1 mg/dL, the hazard in the group with higher LDL is estimated to be 2.88% higher on average.

* 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**. If all of the slope parameters are equal across knots in our Cox splines model, then the linear splines reduce to a straight line. Hence we test the hypothesis for that all the slope parameters in our model from (a) are equal.

**Inference**. The chi-square test of the equality of the slope parameters in our Cox splines model has a p-value of 0.0788. We do not have statistically significant evidence that the slopes across LDL categories are unequal, hence a continuous linear term for LDL would be expected to provide a better fit for the hazard ratio than the linear splines. This test does not rule out the existence of a true association that is non-linear.

* 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.

**Answer**. See graph for 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?

**Answer**. In the first three homeworks we dichotomized cut-offs for survival at 5 years and LDL at 160 mg/dL, and always used at least one of these variables in the models or tests we used to investigate the possible association between all-cause mortality and serum LDL. When we dichotomize a continuous variable in this way, we lose much of the information in the data set. This loss of information can lead to imprecise estimates.

The Cox proportional hazards regression models in Homeworks 4 and 5 considered survival time as a continuous, right-censored variable and treat LDL as a (possibly transformed) continuous variable, or at least consider more than two groups based on LDL (i.e. the categories based on the Mayo clinic recommendations). Thus we are able to use more information from our sample to make inferences about possible associations.

* 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.

**Answer**. The fitted values from all five Cox models we have considered are presented in the graph below. In Homework 4, we fit Cox regressions using LDL, log-LDL, and quadratic-LDL and saw that in each case there was a generally decreasing trend in the hazard for the models. Additionally, the fitted values for each of the three transformations appeared quite similar for the mid-LDL values. The quadratic-fit model predicted a U-shaped trend, though this was inevitable due to the parabolic nature of the predictors.

In problem 2 of this homework we fit a Cox regression using dummy variables for the LDL categories suggested by the Mayo Clinic. In problem 3, a Cox regression was fit using linear splines separated at knots determined again by the Mayo Clinic cut-offs (i.e. at serum LDL of 70, 100, 130, 160, and 190 mg/dL). The predicted values for these two models again show a generally decreasing trend in hazard as LDL increases to the reference group at 160 mg/dL serum LDL. At approximately 190 mg/dL, we see both of these models predicting an increase in the relative hazard. This U-shape in the fitted values was not guaranteed (unlike the quadratic fit from HW4) by the model we chose, though we saw earlier in this HW that we do not have statistically significant evidence that these models provide a better fit than the model using linear continuous LDL as a predictor.



* 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?

**Answer**. In exploring the association between all-cause mortality and serum LDL in an elderly population, the method we settle on a priori should take advantage of the continuous measurements of serum LDL and the knowledge of right censoring in the survival data. For ease of scientific interpretation of the coefficients in our model, fitting a Cox proportional hazards regression on log-LDL seems most reasonable. The logarithm makes sense here since LDL can be expected to have a high level of variance in an elderly population, and the transformation may help stabilize the variances of our estimates in this case. To avoid unnecessary distributional assumptions, we should also use robust standard errors.