**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: To examine the association between odds of diabetes and race, a logistic regression model was fit to examine the outcome of diabetes as the odds ratio (OR) of diabetes between predictors groups defined by race, compared to the referent racial group “white”. The outcome is presence of diabetes, which is a binary outcome. Odds of diabetes in a given racial group and OR of diabetes in the given predictor racial group is compared to the referent racial group “white” with the logistic regression model. Statistical inference was based on the Wald Chi-squared statistic computed using robust estimation where we do not assume equal variance. This is a saturated model as race has 4 categories and we have 4 model parameters (3 slopes and 1 intercept).

Inference: The study consisted of 735 enrolled subjects, who were evaluated for the presence of diabetes prior to receiving an MRI. There was no missing data for prevalence of diabetes or race. The distribution of race showed that there were 572 whites, 104 blacks, 47 Asians and 12 subjects of other race. Diabetes was diagnosed in 79 patients in total. Using logistic regression to model the OR for diabetes in blacks compared to whites is 1.929, the OR for diabetes in Asians compared to whites is 0.6282, and the OR for diabetes in those of other race compared to whites is 1.843. We will not assess statistical significance or confidence intervals between individual groups, but rather by assessing the overall Chi-squared statistic, we see that we cannot reject the null hypothesis that the odds of diabetes is the same between groups (p = 0.0956).

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

β0 (intercept) = -2.221 is the log odds of diabetes when the race is white.

e^β0 = 0.1085 is the odds of diabetes when the race is white.

β1 = 0 .6568 is the difference in the log odds of diabetes when the race is black compared to white (i.e. log odds (DM | black) – log odds (DM | white).

e^β1 = 1.929 is the OR of diabetes when the race is black compared to white (i.e. odds (DM | black) / odds (DM | white).

β2 = -0 .4648 is the difference in the log odds of diabetes when the race is Asian compared to white (i.e. log odds (DM | Asian) – log odds (DM | white).

e^β2 = 0.6282 is the OR of diabetes when the race is Asian compared to white (i.e. odds (DM | Asian) / odds (DM | white).

β3 = 0.6113 is the difference in the log odds of diabetes when the race is other compared to white (i.e. log odds (DM | other) – log odds (DM | white).

e^β3 = 1.843 is the OR of diabetes when the race is other compared to white (i.e. odds (DM | other) / odds (DM | white).

* 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.

Ignoring issues related to multiple comparisons, we would conclude that there is a statistically significant association between odds of diabetes and race only for blacks and whites (p=0.026), but not between Asians & whites (p=0.449) or other races and whites (p=0.438). However, looking at individual p-values introduces multiple comparisons, so is not valid to use these p-values. Since these dummy variables are actually mutually exclusive categories of a single variable, they can be best compared with a much more simple procedure, such as the chi-squared test, which predicts that we cannot reject the null hypothesis that the distribution of the odds of diabetes across racial groups is the same (p = 0.096).

* 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: To examine the association between prevalence of diabetes and race, a logistic regression model was fit to examine the outcome of diabetes as the odds ratio (OR) of diabetes between predictors groups defined by race, compared to the referent racial group “black”. The outcome is presence of diabetes, which is a binary outcome. Odds ratio of diabetes in the given predictor racial group compared to the referent racial group “black”. Statistical inference was based on the Wald Chi-squared statistic computed using robust estimation where we do not assume equal variance. This is a saturated model as race has 4 categories and we have 4 model parameters (3 slopes and 1 intercept).

Inference: The study consisted of 735 enrolled subjects, who were evaluated for the presence of diabetes prior to receiving an MRI. There was no missing data for prevalence of diabetes or race. The distribution of race showed that there were 572 whites, 104 blacks, 47 Asians and 12 subjects of other race. Diabetes was diagnosed in 79 patients in total. Using logistic regression to model the OR for diabetes in whites compared to blacks is 0 .5185, the OR for diabetes in Asians compared to whites is 0 .3258, and the OR for diabetes in those of other race compared to whites is 0.9556. We will not assess statistical significance or confidence intervals between individual groups, but rather by assessing the overall Chi-squared test statistic, we see that we cannot reject the null hypothesis that the odds of diabetes is the same between groups (p = 0.0956).

This regression model uses “black” as the referent racial category in computing ORs for diabetes, whereas the regression model in part (a) uses “white” as the referent racial category in computing ORs for diabetes. The report of formal inference changes with this reparameterization in that the referent group is different, so the ORs are different for Asians and other, whereas the OR for black white is the inverse. The Chi-squared statistic does not change and our overall assessment is that we cannot reject the null hypothesis that the odds of diabetes are the same between racial groups.

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

β0 (intercept) = -1.564 is the log odds of diabetes when the race is black.

e^β0 = 0.2093 is the odds of diabetes when the race is black.

β1 = -0 .6568 is the difference in the log odds of diabetes when the race is white compared to black (i.e. log odds (DM | white) – log odds (DM | black).

e^β1 = 0.5185 is the OR of diabetes when the race is white compared to black (i.e. odds (DM | white) / odds (DM | black).

β2 = -1.122 is the difference in the log odds of diabetes when the race is Asian compared to black (i.e. log odds (DM | Asian) – log odds (DM | black).

e^β2 = 0.3258 is the OR of diabetes when the race is Asian compared to white (i.e. odds (DM | Asian) / odds (DM | black).

β3 = -0.0455 is the difference in the log odds of diabetes when the race is other compared to white (i.e. log odds (DM | other) – log odds (DM | black).

e^β3 = 0.9556 is the OR of diabetes when the race is other compared to white (i.e. odds (DM | other) / odds (DM | black).

* 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.

Ignoring issues related to multiple comparisons, we would conclude that there is a statistically significant association between odds of diabetes and race only for white and black (p=0.026), but not between Asian & black (p=0.085) or those of other race and black (p=0.956). However, looking at individual p-values introduces multiple comparisons, so is not valid to use these p-values. Since these dummy variables are actually mutually exclusive categories of a single variable, they can be best compared with a much more simple procedure, such as the chi-squared test, which predicts that we cannot reject the null hypothesis that the distribution of the odds of diabetes across racial groups is the same (p = 0.096).

* 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)?

Since it is arbitrary which referent group we choose, comparing p-values for individual regression parameters from a dummy variable regression is not appropriate to include or exclude those variables in a multiple regression model (i.e. in a “stepwise model building” procedure). This is because the dummy variables are mutually exclusive categories of a single variable; they can be best compared with a much more simple procedure, such as the chi-squared test. Using the individual p-values from each regression parameter introduces the problem of multiple comparisons where we are more likely to find an association when one does not exist (type I error).

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.

Statistical methods for descriptive statistics: The censoring distribution was described for the population who had LDL baseline measurement, as well as separately for those with baseline LDL who died or were otherwise censored using Kaplan-Meier estimates of the average time to death or censoring, the 50th (median) percentiles and range. Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the range, mean, standard deviation, and the 25th, 50th (median), 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: < 70mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and ≥ 190mg/dL. Within these categories, Kaplan-Meier estimates of survival were calculated and graphed. Estimates by Kaplan Meier of the 1, 3 and 5 year survival probabilities, and the 10th and 20th percentiles of survival distribution were given. Mean survival as well as restricted mean survival during a period of observation that all survivors were followed (5.0 years) are also presented.

Descriptive statistics: The study consisted of 735 subjects who were followed for death from any cause. 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, the other 8 were censored after 5 years. No patient was followed up for less than 5 years unless that patient died. Of the 725 subjects with baseline serum LDL measurement, the Kaplan-Meier (K-M) estimated average time of follow-up for death from any cause was 4.94 years (median 5.14 years, range: 0.19 – 5.91 years), during which time 131 deaths were observed. Of these 131 deaths, the K-M average time of follow-up was 3.23 years (median 3.47 years, range: 0.19 – 5.54 years). Of the 594 study survivors, the K-M average time of follow-up was 5.32 years (median 5.18 years, range: 5.00 – 5.91 years). 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). The mean LDL was 127.4 mg/dL in the 594 survivors and 118.6 in the 131 persons who died during the study. Table 1 presents estimates of the survival distribution within strata defined by serum LDL (by the Mayo Clinic cutoffs), in the combined sample from the 725 subjects with available LDL measurements and in the 10 subjects without baseline measurement of LDL. 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 (88.0%). On average, the subjects in the lowest LDL stratum were estimated to average 4.91 years of life while the other strata averaged from 5.33 to 5.58 years. When restricting the analysis to those who survived for at least 5 years, the subjects with the lowest LDL stratum were estimated to average 4.95 years of life while the other strata averaged from 5.33 to 5.60 years.

**Table 1: Kaplan-Meier based estimates of survival probability and distribution of time from study enrollment to death from any cause, by categories of baseline serum LDL level, given by Mayo Clinic cutoffs for LDL.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Serum LDL at Study enrollment (by Mayo Clinic cutoffs): | Subjectswith LDLAvailable | Subjects with LDL Unavailable |
| Ideal if high risk\*: 11-69 mg/dL | Ideal if at risk\*: 70 – 99mg/dL | Near ideal: 100 – 129mg/dL | Borderline high: 130 – 159mg/dL | High: 160 – 189mg/dL | Very high:190 – 247mg/dL |
| Subjects, n | 22 | 143 | 228 | 225  | 83  | 24  | 725 | 10 |
| Deaths, n | 10  | 28  | 44  | 34  | 11  | 4  | 131 | 2 |
| SurvivalProbability, - 1 year- 3 year- 5 year | 1.000.9090.591 | 0.9790.9090.832 | 0.9830.9120.811 | 0.9780.9290.871 | 1.000.9640.880 | 1.000.9170.833 | 0.9830.9230.836 | 0.8000.8000.800 |
| Percentile of Survival (yrs)- 90th - 80th  | 3.463.55 | 3.805.44 | 3.415.36 | 4.29N/A | 4.53N/A | 4.12N/A | 3.665.54 | 1.890.66 |

**Figure 1: Kaplan-Meier survival probability, by categories of baseline serum LDL level as given by the Mayo Clinic cutoffs for LDL.**

****

Statistical methods for inferential statistics: To examine the association between all-cause mortality and serum LDL, the 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 when fit as dummy variables using the categories suggested by the Mayo Clinic (as above). Quantification of association between all-cause mortality was summarized by the hazards ratio computed from the regression model, with two-sided p values computed using Wald Chi-squared statistic. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.

Statistical Inference: Descriptive statistics are given above. Using proportional hazards regression model, the HR for death in those subject with baseline LDL from 70 – 99 mg/dL was 0.3980 when compared to the group with serum LDL ≤ 70. The HR for death in those subjects with baseline LDL from 100 – 129 mg/dL was 0.3926 when compared to the group with serum LDL ≤ 70. The HR for death in those subjects with baseline LDL from 130 – 159 mg/dL was 0.2939 when compared to the group with serum LDL ≤ 70. The HR for death in those subjects with baseline LDL from 160 – 189 mg/dL was 0.2565 when compared to the group with serum LDL ≤ 70. The HR for death in those subjects with baseline LDL ≥ 190 mg/dL was 0.3167 when compared to the group with serum LDL ≤ 70. We will not assess statistical significance or confidence intervals between individual groups, but rather by assessing the overall Chi-squared statistic, we see that with high statistical significance (p = 0.0087) we can reject the null hypothesis that the survival experience was the same between groups, in favor of a worse survival for those in the lowest serum LDL category (LDL ≤ 70 mg/dL).

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

β0 (intercept) is the log hazard of death when the LDL < 70 mg/dL [not given in the regression output].

e^β0 is the hazard of death when the LDL < 70 mg/dL [not given in the regression output].

β1 is the difference in the log hazard of death when the LDL is 70 - 99 mg/dL compared to when the LDL is < 70 mg/dL (i.e. log hazard (death | ldl70) – log hazard (death | ldl0).

e^β1 = 0.3980 is the HR of death when the LDL is 70 – 99 mg/dL compared to LDL < 70 mg/dL (i.e. hazard (death | LDL 70 – 99) / hazard (death | LDL < 70).

β2 is the difference in the log hazard of death when the LDL is 100 - 129 mg/dL compared to when the LDL is < 70 mg/dL (i.e. log hazard (death | ldl100) – log hazard (death | ldl0).

e^β2 = 0.3926 is the HR of death when the LDL is 100 – 129 mg/dL compared to LDL < 70 mg/dL (i.e. hazard (death | LDL 100 - 129) / hazard (death | LDL < 70).

β3 is the difference in the log hazard of death when the LDL is 130 - 159 mg/dL compared to when the LDL is < 70 mg/dL (i.e. log hazard (death | ldl130) – log hazard (death | ldl0).

e^β3 = 0.2939 is the HR of death when the LDL is 130 – 159 mg/dL compared to LDL < 70 mg/dL (i.e. hazard (death | LDL 130 – 159) / hazard (death | LDL < 70).

β4 is the difference in the log hazard of death when the LDL is 160 - 189 mg/dL compared to when the LDL is < 70 mg/dL (i.e. log hazard (death | ldl160) – log hazard (death | ldl0).

e^β4 = 0.2565 is the HR of death when the LDL is 160 – 189 mg/dL compared to LDL < 70 mg/dL (i.e. hazard (death | LDL 160 – 189) / hazard (death | LDL < 70).

β5 is the difference in the log hazard of death when the LDL > 190 mg/dL compared to when the LDL is < 70 mg/dL (i.e. log hazard (death | ldl190) – log hazard (death | ldl0).

e^β5 = 0.3167 is the HR of death when the LDL is > 190 mg/dL compared to LDL < 70 mg/dL (i.e. hazard (death | LDL > 190mg/dL) / hazard (death | LDL < 70).

* 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?

To assess whether the regression model for hazard ratio for death given LDL with LDL modelled with dummy variables is a “better fit” than does a model that uses only a continuous linear term for LDL, I would put a linear term into the model as assess if the slopes are equal to zero using post-estimation commands. We should reject the null hypothesis that the model based on dummy variables of categorization by the May Clinic cutoffs is different from a linear model (p = 0.0089), i.e. we favor the alternative hypothesis that the model is non-linear.

We should reject the null hypothesis that the model based on splines of LDL using the Mayo Clinic cutoffs is the same as a linear model (p< 0.0001), i.e. we favor the alternative hypothesis that the model 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.
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: See above for descriptive statistics. To examine the association between all-cause mortality and serum LDL, the 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 with splines using the categories suggested by the Mayo Clinic (as above). Quantification of association between all-cause mortality was summarized by the hazards ratio computed from the regression model, for each 10 point rise in serum LDL within the range of the spline. Inference was made with two-sided p values computed using Wald Chi-squared statistic. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.

Inference: Descriptive statistics are given above. Using the proportional hazards regression model, the HR for death in those subjects with baseline serum LDL < 70 mg/dL is 0.8014 for each 10 mg/dL increase from the lower group. The HR for death in those subjects with baseline serum LDL 70 – 99 mg/dL is 0.8148 for each 10 mg/dL increase from the lower group. The HR for death in those subjects with baseline serum LDL 100 – 129 mg/dL is 0.9773 for each 10 mg/dL increase from the lower group. The HR for death in those subjects with baseline serum LDL 130 – 159 mg/dL is 1.0367 for each 10 mg/dL increase from the lower group. The HR for death in those subjects with baseline serum LDL 160 – 189 mg/dL is 0.7444 for each 10 mg/dL increase from the lower group. The HR for death in those subjects with baseline serum LDL > 190 mg/dL is 1.3284 for each 10 mg/dL increase from the lower group.

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

β1 is the difference in the log hazard of death for each 1 mg/dL increase in serum LDL when the LDL is < 70 mg/dL (i.e. log hazard (death | ldlx+1) – log hazard (death | ldlx), for LDL < 70).

β2 is the difference in the log hazard of death for each 1 mg/dL increase in serum LDL when the LDL is 70 – 99 mg/dL (i.e. log hazard (death | ldlx+1) – log hazard (death | ldlx), for LDL from 70 - 99).

β3 is the difference in the log hazard of death for each 1 mg/dL increase in serum LDL when the LDL is 100 – 129 mg/dL (i.e. log hazard (death | ldlx+1) – log hazard (death | ldlx), for LDL from 100 - 129).

β4 is the difference in the log hazard of death for each 1 mg/dL increase in serum LDL when the LDL is 130 – 159 mg/dL (i.e. log hazard (death | ldlx+1) – log hazard (death | ldlx), for LDL from 130 - 159).

β5 is the difference in the log hazard of death for each 1 mg/dL increase in serum LDL when the LDL is 160 – 189 mg/dL (i.e. log hazard (death | ldlx+1) – log hazard (death | ldlx), for LDL from 160 - 189).

β6 is the difference in the log hazard of death for each 1 mg/dL increase in serum LDL when the LDL is ≥ 190mg/dL (i.e. log hazard (death | ldlx+1) – log hazard (death | ldlx), for LDL ≥ 190).

e^β1 = 0.9781 is the HR of death for each 1 mg/dL increase in serum LDL, when the LDL < 70 mg/dL.

e^(β1\*10) = (HR)^10 = 0.8014 is the HR of death for each 10 mg/dL increase in serum LDL, when the LDL < 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?

To assess whether the regression model for hazard ratio for death given LDL with LDL modelled with splines based on cutoffs from the Mayo clinic is a “better fit” than does a model that uses only a continuous linear term for LDL, I would put a linear term into the model and assess if the slopes are equal to zero using post-estimation commands. We should reject the null hypothesis that the model based on splines of LDL using the Mayo Clinic cutoffs is the same as a linear model (p< 0.0001), i.e. we favor the alternative hypothesis that the model 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.
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?

The regression strategies in homeworks 4 and 5 use proportional hazards regression, or survival analysis. This is the best approach to use for censored data as we can take into account he amount of follow-up time that each subject contributes. Furthermore, these approaches used LDL as a continuous measure and death as a time-varying outcome, without the need to dichotomize as we did in homeworks 1-3.

* 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.

Figure 2: Plots of fitted hazard ratios from the proportional hazards regression models where LDL is modelled as dummy variables or splines, based on cutoffs suggested by the Mayo Clinic. All hazard ratio estimates are relative to a group having serum LDL of 160mg/dL.



Figure 3: Plots of fitted hazard ratios from the proportional hazards regression models where LDL is modelled as linear continuous, logarithmically transformed, quadratic function (including both linear and squared terms), dummy variables or splines. Inflection points for the latter two models are based on cutoffs suggested by the Mayo Clinic. All hazard ratio estimates are relative to a group having serum LDL of 160mg/dL.



Figure 2 displays the fitted values from the 2 models presented in this homework. In each case, the model predicts a trend which is predominantly towards lower relative hazard with increasing LDL. Neither of these models from this homework fits the continuously modelled LDL linear, logarithmic or quadratic models from homework #4. Again, the fitted relative hazards are most discrepant where our data is sparse, especially at the low end of the range of LDL. These models from this homework are not improvements on the models from last homework.

* 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*, of all the analyses performed thus far to explore for an unadjusted association between all-cause mortality and serum LDL in an elderly population, I prefer the proportional hazards model with LDL modelled as a continuous logarithmically-transformed variable. This model allows for assessment of contribution of person-time to the analysis and is more precise than the PH model with LDL modelled as a continuous linear variable (the latter model is only slightly easier to interpret).

**Discussion Sections: February 3 - 7, 2014**

We continue to discuss the dataset regarding FEV and smoking in children. Come do discussion section prepared to describe descriptive statistics, especially as they relate to confounding, precision, effect modification, and the impact of heteroscedasticity.