Biost 518: Applied Biostatistics II
Biost 515: Biostatistics II
Emerson, Winter 2014

Homework #5 Key
February 10, 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.

***Instructions for grading:** Prior to the answer for each problem, I provide the maximum points to be given for each problem, and the way that points should be distributed. Please insert comments on to the document indicating the points you have awarded for the problem, commenting on any reasons points were deducted.*
***My answer to each question is provided in boldface type. In giving the answers, I sometimes provide alternative approaches in order that you can assess whether the numbers match up. I also provide some discussion of the choices or some additional material that I did not really expect to be provided in the answer. This additional information is provided in** normal type.*

1. Perform a statistical regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across.
    a. 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.

Answer: (10 points) This is a saturated model, because we are fitting four regression parameters (an intercept and 3 slopes) to a datset having four distinct racial groups.

Statistical Methods for inferential statistics: Distributions of prevalence of diabetes was compared across groups defined racial groups (white, black, Asian, other) using a logistic regression model with dummy variables. Quantification of any association between prevalence of diabetes and race was summarized by the three odds ratios (ORs) computed from the regression model, comparing each of the other racial groups to whites. Confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator. No adjustment was made for the multiple comparisons inherent in the comparisons within each of the paired comparisons, and thus the confidence intervals and p values quantifying associations within those strata should be judged as purely descriptive. No subjects were missing data for either variable. *(Note my commenting on the lack of adjustment for multiple comparisons.This is generally the way I handle the reporting of dummy variables: the reader is just as capable of applying a Bonferroni correction as I am, and I did not do anything fancier. I could, of course, have used some other stratum as my reference group.)*

Inferential results: Data was available on 735 subjects, of whom 572 were white, 104 were black, 44 were Asian, and 12 were other. A total of 79 (10.9%) had prevalent diabetes. The prevalence within racial groups was 9.79% (56 of 572) in whites, 17.31% (18 of 104) in blacks, 6.38% (3 of 44) in Asians, and 16.7% (2 of 12) in those of other races. The logistic regression model estimates the odds of prevalent diabetes to be

- **1.93 fold higher in blacks than whites (OR 1.93, unadjusted 95% CI 1.08 – 3.44, P = 0.026 not adjusted for multiple comparisons),**
- **0.628 as high in Asians as whites (OR 0.628, unadjusted CI 0.189 – 2.09, P = 0.448 not adjusted for multiple comparisons), and**
- **1.84 fold higher in person of other races than in whites (OR 1.84, unadjusted CI 0.394 – 8.62, P = 0.437 not adjusted for multiple comparisons).**

Such estimated differences were not statistically significant: A test for inequality among the groups was not statistically significant (P= 0.110).

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

Answer: (3 points)

- **The exponentiated intercept of $e^{-2.221} = 0.109$ corresponds to the odds of having diabetes in whites.**
- **The exponentiated slope of 1.93 labeled race 2 is the odds ratio comparing blacks (numerator) to whites (denominator).**
- **The exponentiated slope of 0.628 labeled race 3 is the odds ratio comparing Asians (numerator) to whites (denominator).**
- **The exponentiated slope of 1.84 labeled race 4 is the odds ratio comparing other racial groups (numerator) to whites (denominator).**

c. 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: (3 points) Using the Wald based p values reported with the regression parameter estimates and a 0.05 level of significance, we would conclude there was a statistically significant difference between blacks and whites with respect to prevalence of diabetes. We would conclude that we did not have sufficient evidence to declare a difference between whites and either Asians or other racial groups. (The regression output gave us no information about other pairwise comparisons.)

d. 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)?

Answer: (3 points) Had I used blacks as the reference group, I would have perhaps descriptively reported the ORs comparing whites, Asians, and other racial groups to blacks, but the conclusion of lack of statistical evidence of an association remains unchanged. This is just a reparameterization of the same model.

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

Answer: (3 points)

- **The exponentiated intercept of $e^{-1.564} = 0.209$ corresponds to the odds of having diabetes in blacks.**
- **The exponentiated slope of 0.519 labeled race 1 is the odds ratio comparing whites (numerator) to blacks (denominator). This is just the reciprocal of 1.93.**
- **The exponentiated slope of 0.326 labeled race 3 is the odds ratio comparing Asians (numerator) to blacks (denominator).**
- **The exponentiated slope of 0.956 labeled race 4 is the odds ratio comparing other racial groups (numerator) to blacks (denominator).**

f. 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: (3 points) Using the Wald based p values reported with the regression parameter estimates and a 0.05 level of significance, we would conclude there was a statistically significant difference between blacks and whites with respect to prevalence of diabetes. We would conclude that we did not have sufficient evidence to declare a difference between blacks and either Asians or other racial groups. (The regression output gave us no information about other pairwise comparisons.)

g. 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: (5 points) If we were to foolishly delete regression tems that were not significant, in the first model we would drop the terms for Asians and other races, thus treating them in the same category as whites. In the second model, we would again drop the terms for Asians and other races, now treating them in the same category as blacks. Hence, we would end up with very different inference that would be due solely to an extremely arbitrary choice of how to code the race variable. This is not good. Don't do it. When trying to make inference about "variable importance", nclude all terms for a dummy variable modeling, or include none of them.

2. 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)
```

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

Answer: (10 points) *(The following is copied directly from my Homework #4 Key).*

Statistical Methods for descriptive statistics: Descriptive statistics for the censoring distribution included the minimum and maximum observed censoring times and the Kaplan-Meier estimates of the 10^{th}, 50^{th} (median), and 90^{th} percentiles, as well as the mean time of follow-up calculated as the area under the Kaplan-Meier estimate of the censoring distribution's survivor curve.

Descriptive statistics for serum LDL levels included the number of cases with missing data, as well as the minimum, maximum, mean, standard deviation, and the 25^{th}, 50^{th} (median), and 75^{th} percentiles for the cases with available data. For the purposes of descriptive statistics of the survival probabilities by serum LDL level, serum LDL was categorized according to the Mayo Clinic guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Within these categories, Kaplan-Meier estimates of survival were calculated and graphed, and estimates of the 2 and 5 year survival probabilities, as well as the 10^{th} and 20^{th} percentiles of the survival distribution and the restricted mean survival during a period of observation that all LDL strata still had some subjects at risk (5.75 years). *(These descriptive statistics will suffice for all problems. The key issue is that I would want to have multiple strata in order to be able to glean some information about nonlinearity in the association between all cause mortality and LDL. While it is true that the way I might ideally categorize LDL to investigate linearity of log LDL versus the categories I would investigate linearity of untransformed LDL or quadratic relationships, the range of LDL measurements did not really allow substantially different categorizations. Hence, I used the scientifically determined categories based on the Mayo Clinic recommendations.)*

Descriptive statistics: The study consisted of 735 subjects who were followed for death from any cause for a Kaplan-Meier estimated average of 5.33 years (median 5.66 years, range 5.00 to 5.91 years), during which time 133 deaths were observed. Serum LDL measurements at the time of study enrollment were not available on 10 subjects, two of whom were observed to die after 0.189 and 0.657 years of observation, with the remaining subjects still alive after 5.05 to 5.91 years of observation. In the 725 subjects with available serum LDL measurements at enrollment, the mean LDL was 126 mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL).

Table 1 presents estimates of the survival distribution within strata defined by serum LDL and in the combined sample from the 725 subjects with available LDL measurements. The greatest difference in survival distributions is apparent when comparing those individuals having the lowest serum LDL levels (less than 70 mg/dL) at times after 2 years of follow-up. The 5 year survival probability is lowest in that group (59.1%) and is observed highest in the subjects having serum LDL between 160 and 189 mg/dL inclusive (88.0%). On average, the subjects in the lowest LDL stratum were estimated to average 4.91 years of life during the first 5.75 years following study enrollment, while the other strata averaged from 5.23 to 5.45 years. Figure 1 presents the Kaplan-Meier survival probability estimates graphically, where it is again the lowest LDL group that shows the most markedly different survival distribution.

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

	Serum LDL at Study Enrollment						All Subjects (with LDL Available[3])
	11 – 69 mg/dL	70 – 99 mg/dL	100 – 129 mg/dL	130 – 159 mg/dL	160 – 189 mg/dL	190 – 247 mg/dL	
N Subjects	22	143	228	225	83	24	725
N Deaths	10	28	44	34	11	4	131
2 year Survival Probability[1]	100%	95.8%	93.9%	95.6%	98.8%	95.8%	95.6%
5 Year Survival Probability[1]	59.1%	83.2%	81.1%	87.1%	88.0%	83.3%	83.6%
10th Pctile of Survival[1]	3.46 y	3.80 y	3.41 y	4.30 y	4.53 y	4.13 y	3.66 y
20th Pctile of Survival[1]	3.55 y	5.44 y	5.36 y	NA[1]	NA[1]	NA[1]	5.54 y
5.75 Year Restricted Mean of Survival[2]	4.91 y	5.24 y	5.23 y	5.35 y	5.45 y	5.32 y	5.29 y

[1] Based on Kaplan-Meier estimates computed within strata defined by LDL and overall. NA indicates that the corresponding percentile is not estimable with the available data.

[2] Average number of years alive during the first 5.75 years following study enrollment, as computed by the area under Kaplam-Meier survival curves computed within strata defined by LDL and overall

[3] Ten of the 735 subjects in the study population were missing baseline serum LDL measurements. Two of those subjects were observed to die after 0.189 y and 0.657 years of observation. The remaining 8 subjects with missing LDL data were still alive at the end of their observation period 5.03 to 5.91 years after study enrollment

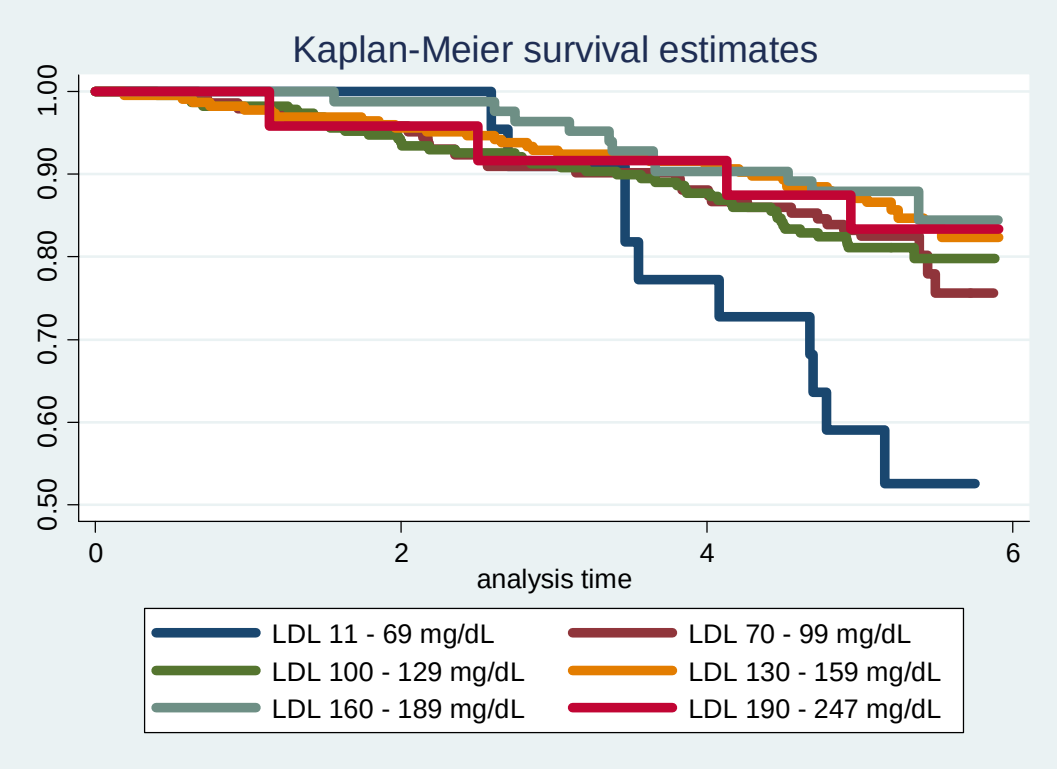


Figure 1: Kaplan-Meier based estimates of distribution of time from study enrollment to death from any cause for 725 subjects having serum LDL measurements at baseline.

<u>Statistical Methods for inferential statistics</u>: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as dummy variables over the intervals 11 – 69 mg/dL, 70 – 99 mg/dL, 100 – 129 mg/dL, 130 – 159 mg/dL, 160 – 189 mg/dL, and 190 – 247 mg/dL Quantification of any association between all cause mortality and serum LDL was summarized by the five hazards ratios (HRs) computed from the regression model, comparing each of the higher LDL groups to the group having LDL less than 70 mg/dL. Confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator. No adjustment was made for the multiple comparisons inherent in the comparisons of each of the five higher LDL strata to the lowest stratum, and thus the confidence intervals and p values comparing those strata should be judged as purely descriptive. The linearity of association between serum LDL and the log hazard function was effected using a model that included both the linear continuous term and all of the above described dummy variables. A hypothesis test used a Wald statistic to test that the regression parameter estimates for the dummy variables would all be 0. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis. *(Note my commenting on the lack of adjustment for multiple comparisons.This is generally the way I handle the reporting of dummy variables: the reader is just as capable of applying a Bonferroni correction as I am, and I did not do anything fancier. I could, of course, have used some other stratum as my reference group.)*

<u>Inferential results</u>: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis fitting dummy variables to the strata used in Mayo Clinic recommendations, we find evidence of a statistically significant association between the instantaneous risk of death from all causes and serum LDL levels (P = 0.0087). From that regression analysis, we estimate that compared to a population having serum LDL less than 70 mg/dL, the instantaneous risk of death is

- **a relative 60.2% lower in the group having LDL levels between 70 and 99 mg/dL (HR = 0.398, with 95% CI 0.203 – 0.782, P = 0.008 (neither CI nor P value adjusted for multiple comparisons)).**
- **a relative 60.7% lower in the group having LDL levels between 100 and 129 mg/dL (HR = 0.393, with 95% CI 0.207 – 0.744, P = 0.004 (neither CI nor P value adjusted for multiple comparisons)).**
- **a relative 70.6% lower in the group having LDL levels between 130 and 159 mg/dL (HR = 0.294, with 95% CI 0.152 – 0.568, P < 0.0005 (neither CI nor P value adjusted for multiple comparisons)).**
- **a relative 74.3% lower in the group having LDL levels between 160 and 189 mg/dL (HR = 0.257, with 95% CI 0.113 – 0.580, P = 0.001 (neither CI nor P value adjusted for multiple comparisons)).**
- **a relative 68.3% lower in the group having LDL levels greater than or equal to 190 mg/dL (HR = 0.317, with 95% CI 0.101 – 0.989, P = 0.048 (neither CI nor P value adjusted for multiple comparisons)).**

A secondary test for nonlinearity of the association between serum LDL and the log instantaneous risk of death from all causes found no strong evidence that the association was nonlinear (P = 0.399). *(Many articles would just report the overall test of an association in the text, and present the individual stratum specific estimates in a table such as follows:*

	Hazard Ratio (95% CI; P value)[1]
LDL < 70 mg/dL	*1.00 (reference)*
70 mg/dL ≤ LDL < 100 mg/dL	*0.398 (95% CI 0.203 – 0.782; P = 0.008)*
100 mg/dL ≤ LDL < 130 mg/dL	*0.393 (95% CI 0.207 – 0.744; P = 0.004)*
130 mg/dL ≤ LDL < 160 mg/dL	*0.294 (95% CI 0.152 – 0.568; P < 0.0005)*
160 mg/dL ≤ LDL < 190 mg/dL	*0.257 (95% CI 0.113 – 0.580; P = 0.001)*
190 mg/dL ≤ LDL	*0.317 (95% CI 0.101 – 0.989; P = 0.048)*

[1] *Confidence intervals and p values are not adjusted for the multiple comparisons inherent in the statistical model. Overall test of an association is highly statistically significant (P = 0.0087)*

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

Answer: (5 points) The "intercept" in a proportional hazards model is the hazard function in the "reference group". In this model, the baseline hazard function corresponds to the group having serum LDL less than 70 mg/dL. As described above, the slope parameters then represent comparisons of each stratum to that baseline group:

- **0.398 is the hazard ratio comparing a group having LDL levels between 70 and 99 mg/dL to a group having LDL levels less than 70 mg/dL.**
- **0.393 is the hazard ratio comparing a group having LDL levels between 100 and 129 mg/dL to a group having LDL levels less than 70 mg/dL.**
- **0.294 is the hazard ratio comparing a group having LDL levels between 130 and 159 mg/dL to a group having LDL levels less than 70 mg/dL.**
- **0.257 is the hazard ratio comparing a group having LDL levels between 160 and 189 mg/dL to a group having LDL levels less than 70 mg/dL.**

- **0.317 is the hazard ratio comparing a group having LDL levels greater than or equal to 190 mg/dL to a group having LDL levels less than 70 mg/dL**

c. 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?

Answer: (5 points) As described above, I included a linear continuous term along with the dummy variables, and then tested that the regression coefficients for the dummy variables were 0 in a multiple partial Wald test having a chi square distribution with 5 degrees of freedom. The P value of 0.399 suggests that we do not have sufficient evidence of a trend in the log hazard that is nonlinear, when we use the dummy variables to model departures from nonlinearity. *(Note that when using dummy variables based on categorization of a continuous variable, the straight line model is not a special case. Hence, I had to add a linear term to the model.)*

d. 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: *(No answer necessary. The desired fitted values are displayed in problem 4.)*

3. 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
```

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

Answer: (10 points) Statistical methods and results for descriptive statistical analysis are presented in problem 2.

Statistical Methods for inferential statistics: Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as linear splines over the intervals 11 – 69 mg/dL, 70 – 99 mg/dL, 100 – 129 mg/dL, 130 – 159 mg/dL, 160 – 189 mg/dL, and 190 – 247 mg/dL Quantification of any association between all cause mortality and serum LDL was summarized by the six hazards ratios (HRs) computed from the regression model, comparing two groups within the same stratum that differ by 1 mg/dL in their LDL. Confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator. No adjustment was made for the multiple comparisons inherent in the comparisons within each of the six LDL strata, and thus the confidence intervals and p values quantifying associations within those strata should be judged as purely descriptive. The linearity of association between serum LDL and the log hazard function was effected using this same model by testing that all regression coefficients were equal using a Wald statistic. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis. *(Note my commenting on the lack of adjustment for multiple comparisons.This is generally the way I handle the reporting of dummy variables: the reader is just as capable of applying a Bonferroni correction as I am, and I did not do anything fancier. I could, of course, have used some other stratum as my reference group.)*

Inferential results: Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis fitting linear splines to the strata used in Mayo Clinic recommendations, we find evidence of a statistically significant association

between the instantaneous risk of death from all causes and serum LDL levels (P < 0.0001). From that regression analysis, we estimate that within each of the defined LDL strata:

- **For two groups, both having LDL less than 70 mg/dL we estimate that the group with the higher LDL would have a relative 2.19% lower instantaneous risk of death for each 1 mg/dL difference in the LDL between the two groups (HR = 0.978 per 1 mg/dL higher LDL within the stratum, with 95% CI 0.960 – 0.996, P = 0.019 (neither CI nor P value adjusted for multiple comparisons)).**
- **For two groups, both having LDL between 70 and 100 mg/dL we estimate that the group with the higher LDL would have a relative 2.03% lower instantaneous risk of death for each 1 mg/dL difference in the LDL between the two groups (HR = 0.980 per 1 mg/dL higher LDL within the stratum, with 95% CI 0.953 – 1.01, P = 0.139 (neither CI nor P value adjusted for multiple comparisons)).**
- **For two groups, both having LDL between 100 and 130 mg/dL we estimate that the group with the higher LDL would have a relative 0.229% lower instantaneous risk of death for each 1 mg/dL difference in the LDL between the two groups (HR = 0.998 per 1 mg/dL higher LDL within the stratum, with 95% CI 0.976 – 1.02, P = 0.835 (neither CI nor P value adjusted for multiple comparisons)).**
- **For two groups, both having LDL between 130 and 160 mg/dL we estimate that the group with the higher LDL would have a relative 0.361% higher instantaneous risk of death for each 1 mg/dL difference in the LDL between the two groups (HR = 1.004 per 1 mg/dL higher LDL within the stratum, with 95% CI 0.979 – 1.03, P = 0.773 (neither CI nor P value adjusted for multiple comparisons)).**
- **For two groups, both having LDL between 160 and 190 mg/dL we estimate that the group with the higher LDL would have a relative 2.91% lower instantaneous risk of death for each 1 mg/dL difference in the LDL between the two groups (HR = 0.971 per 1 mg/dL higher LDL within the stratum, with 95% CI 0.930 – 1.01, P = 0.181 (neither CI nor P value adjusted for multiple comparisons)).**
- **For two groups, both having LDL greater than or equal to 190 mg/dL we estimate that the group with the higher LDL would have a relative 2.88% higher instantaneous risk of death for each 1 mg/dL difference in the LDL between the two groups (HR = 1.029 per 1 mg/dL higher LDL within the stratum, with 95% CI 0.979 – 1.08, P = 0.261 (neither CI nor P value adjusted for multiple comparisons)).**

Descriptively, the above regression parameter estimates do not suggest a consistently lower hazard of death with higher LDL. However, when using the linear spline model in a secondary test for nonlinearity of the association between serum LDL and the log instantaneous risk of death from all causes found no strong evidence that the association was nonlinear (P = 0.399). That is, there was not a statistically significant difference among the stratum specific hazard ratios. *(Many articles would just report the overall test of an association in the text, and present the individual stratum specific estimates in a table such as follows:*

	Hazard Ratio per 1mg/dL difference in LDL within the individual strata (95% CI; P value)[1]
LDL < 70 mg/dL	*0.978 (95% CI 0.960 – 0.996; P = 0.019)*
70 mg/dL ≤ LDL ≤ 100 mg/dL	*0.980 (95% CI 0.953 – 1.01; P = 0.139)*
100 mg/dL ≤ LDL ≤ 130 mg/dL	*0.998 (95% CI 0.976 – 1.02; P = 0.835)*
130 mg/dL ≤ LDL ≤ 160 mg/dL	*1.004 (95% CI 0.979 – 1.03; P = 0.773)*

160 mg/dL ≤ LDL ≤ 190 mg/dL	*0.971 (95% CI 0.930 – 1.01; P = 0.181)*
190 mg/dL ≤ LDL	*1.029 (95% CI 0.979 – 1.08; P = 0.261)*

[1] *Confidence intervals and p values are not adjusted for the multiple comparisons inherent in the statistical model. Overall test of an association is highly statistically significant (P = 0.0087)*

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

Answer: (5 points) The "intercept" in a proportional hazards model is the hazard function in the "reference group". In this model, the baseline hazard function corresponds to the group having serum LDL equal to 0 mg/dL. That is outside our data, and not really possible. So the baseline hazard group is not really of interest. As described above, the slope parameters then represent comparisons of each stratum to that baseline group:

- **0.978 is the hazard ratio per 1 mg/dL higher LDL when comparing two groups both having LDL levels less than 70 mg/dL.**
- **0.980 is the hazard ratio per 1 mg/dL higher LDL when comparing two groups both having LDL levels between 70 and 100 mg/dL.**
- **0.998 is the hazard ratio per 1 mg/dL higher LDL when comparing two groups both having LDL levels between 100 and 130 mg/dL.**
- **1.004 is the hazard ratio per 1 mg/dL higher LDL when comparing two groups both having LDL levels between 130 and 160 mg/dL.**
- **0.971 is the hazard ratio per 1 mg/dL higher LDL when comparing two groups both having LDL levels between 160 and 190 mg/dL.**
- **1.029 is the hazard ratio per 1 mg/dL higher LDL when comparing two groups both having LDL levels greater than or equal to 190 mg/dL.**

(Note how I did not need to worry about the endpoints of the intervals when using linear splines. Because the linear splits fit a continuous function, the endpoints can be regarded as part of both adjacent intervals.)

c. 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?

Answer: (5 points) As described above, I used a multiple partial Wald test having a chi square distribution with 5 degrees of freedom to test for equality of all of regression slope coefficients. The P value of 0.079 suggests that we do not have sufficient evidence of a trend in the log hazard that is nonlinear, when we use the linear splines to model departures from nonlinearity. *(Note that when using linear splines based on categorization of a continuous variable, the straight line model is a special case. Hence, I did not have to fit another model.)*

d. 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: *(No answer necessary. The desired fitted values are displayed in problem 4.)*

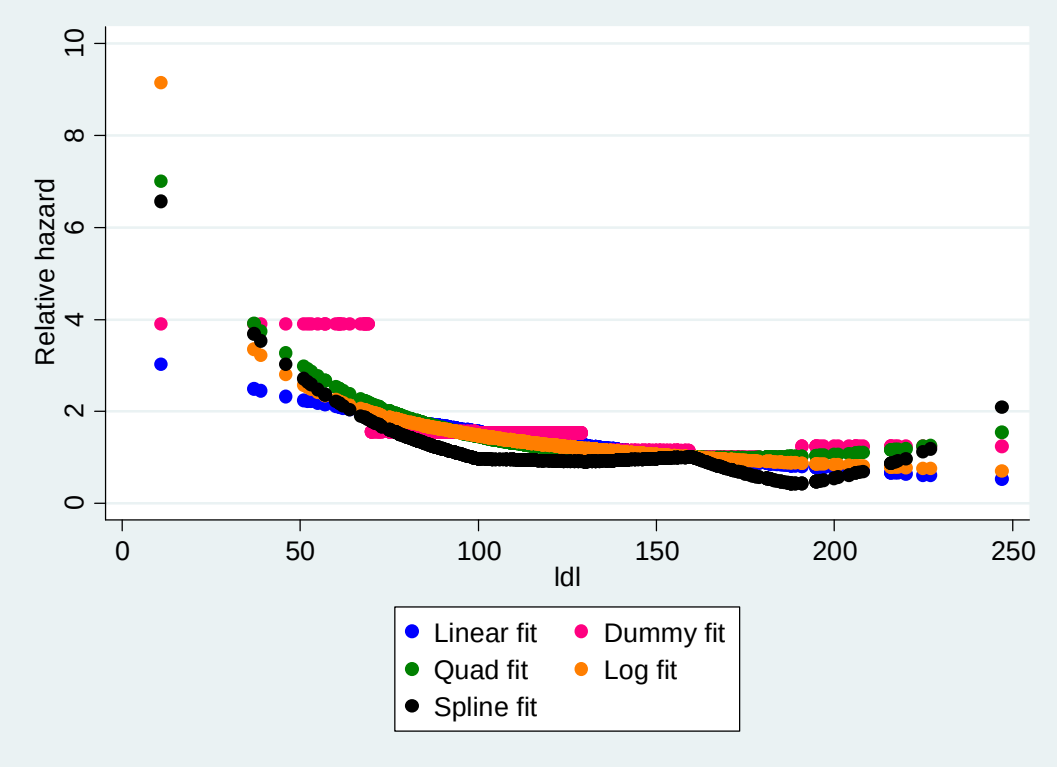
4. By answering the following questions, compare the relative advantages and disadvantages of the various statistical analysis strategies we have considered in Homeworks 1-4 and problems 2 and 3 in this homework.

a. What advantages do the regression strategies used in Homeworks 4 and 5 provide over the approaches used in Homeworks 1-3?

Answer: (3 points) Homeworks 4 and 5 considered the distribution of my putative "outcome" as a continuous variable modeled as a function of groups defined by a continuous function of my putatitive "cause". I find it more pleasing scientifically to use time to death as the response, and anticipate that I avoid the loss of precision that would follow from dichotomization of either LDL or the time to death.

b. 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: (3 points) The following figure displays the fitted HR from all the relevant models. I do not detect much difference between the quadratic, the log transformed, or the linear spline fits. To my eye, the straight line model is much closer to those three than is the dummy variable model. The relevance of this is that the much more flexible models seemed to agree with the log fit as I personally had anticipated, and the dummy variable model seems a very poor choice.



c. *A priori*, of all the analyses we have considered for exploring an (unadjusted) association between all cause mortality and serum LDL in an elderly population, which one would you prefer and why?

Answer: (5 points) When making inference about an association, my top choice a priori would have been the PH regression with the logarithmically transformed LDL. On biological principles, I expected a multiplicative effect across LDL levels such that each doubling of the LDL might be associated with a more constant HR. (I was not so strong in that belief that I would have rejected a linear continuous fit instead, just for the more straightforward interpretation.) But when LDL is my POI, I would have shied away from the more flexible models.

When testing for nonlinearity, I would have used the quadratic as my first choice, but a linear spline model with fewer knots (to avoid unnecessary loss of precision) would also have been acceptable.

When adjusting for LDL as a confounder, I might consider the log fit as likely to be adequate on scientific grounds, but if there was any doubt in my mind I would use a quadratic or possibly a linear spline model with only 3 or 4 knots.

Key among my recommendations is the avoidance of dummy variables, especially when fit to quartiles or quintiles . I have not yet seen a situation where that is superior to one of these other approaches.

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.

Stata Commands and Output

```
. tabulate race diabetes, row col chi

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
+-------------------+

           |       diabetes
      race |         0          1 |     Total
-----------+----------------------+----------
         1 |       516         56 |       572
           |     90.21       9.79 |    100.00
           |     78.66      70.89 |     77.82
-----------+----------------------+----------
         2 |        86         18 |       104
           |     82.69      17.31 |    100.00
           |     13.11      22.78 |     14.15
-----------+----------------------+----------
         3 |        44          3 |        47
           |     93.62       6.38 |    100.00
           |      6.71       3.80 |      6.39
-----------+----------------------+----------
         4 |        10          2 |        12
           |     83.33      16.67 |    100.00
           |      1.52       2.53 |      1.63
-----------+----------------------+----------
     Total |       656         79 |       735
           |     89.25      10.75 |    100.00
           |    100.00     100.00 |    100.00

          Pearson chi2(3) =   6.5836   Pr = 0.086
```

```
.
. logit diabetes i.race

Logistic regression                                     Number of obs   =        735
                                                        LR chi2(3)      =       6.04
                                                        Prob > chi2     =     0.1096
Log likelihood =  -247.7763                             Pseudo R2       =     0.0120

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |
          2  |   .6567795   .2949213     2.23   0.026     .0787444    1.234815
          3  |  -.4648223   .6130707    -0.76   0.448    -1.666419    .7367743
          4  |   .6113172   .7872707     0.78   0.437     -.931705    2.154339
             |
       _cons |  -2.220755   .1406952   -15.78   0.000    -2.496513   -1.944998
------------------------------------------------------------------------------

. logistic diabetes i.race

Logistic regression                                     Number of obs   =        735
                                                        LR chi2(3)      =       6.04
                                                        Prob > chi2     =     0.1096
Log likelihood =  -247.7763                             Pseudo R2       =     0.0120

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |
          2  |   1.928571   .5687768     2.23   0.026     1.081928    3.437742
          3  |   .6282468   .3851597    -0.76   0.448     .1889224    2.089186
          4  |   1.842857   1.450827     0.78   0.437     .3938816    8.622192
------------------------------------------------------------------------------

.
. logit diabetes ib2.race
```

```
Logistic regression                                     Number of obs   =        735
                                                        LR chi2(3)      =       6.04
                                                        Prob > chi2     =     0.1096
Log likelihood =  -247.7763                             Pseudo R2       =     0.0120

------------------------------------------------------------------------------
    diabetes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |
          1  |  -.6567795   .2949213    -2.23   0.026    -1.234815   -.0787444
          3  |  -1.121602   .6505721    -1.72   0.085      -2.3967    .1534961
          4  |  -.0454624    .816813    -0.06   0.956    -1.646386    1.555462
             |
       _cons |  -1.563976   .2591977    -6.03   0.000    -2.071994   -1.055957
------------------------------------------------------------------------------
```

```
. logistic diabetes ib2.race
```

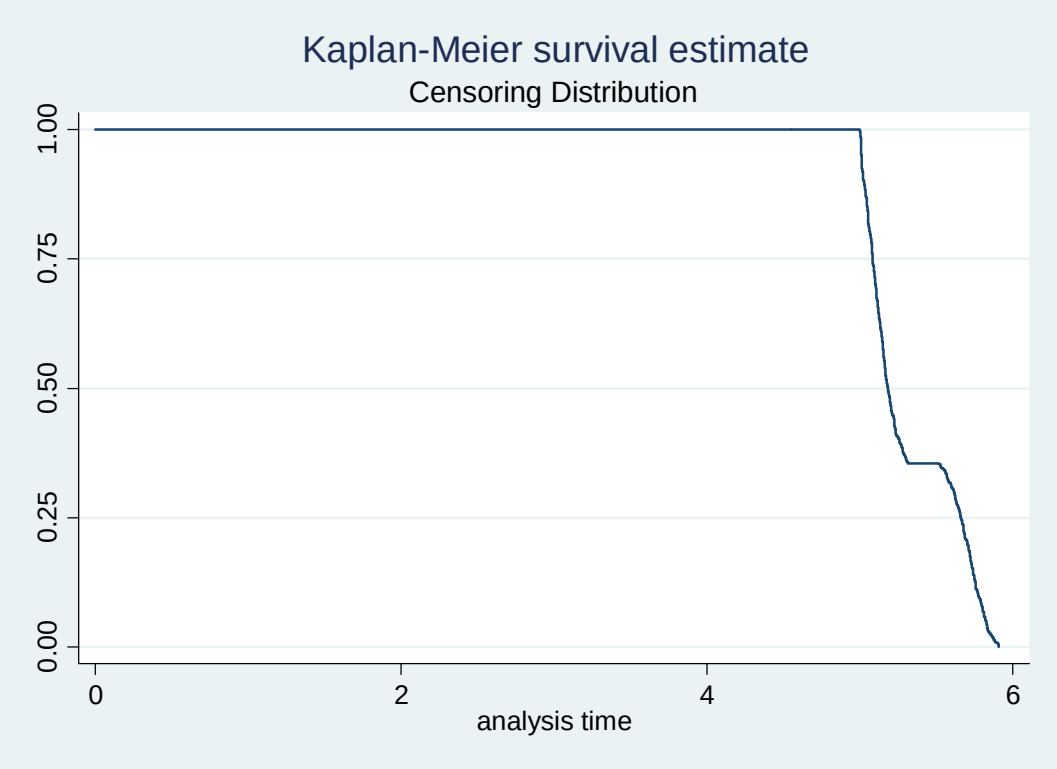
```
Logistic regression                                     Number of obs   =        735
                                                        LR chi2(3)      =       6.04
                                                        Prob > chi2     =     0.1096
Log likelihood =  -247.7763                             Pseudo R2       =     0.0120

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |
          1  |   .5185185   .1529222    -2.23   0.026     .2908887    .9242762
          3  |   .3257576   .2119288    -1.72   0.085     .0910178    1.165903
          4  |   .9555556   .7805102    -0.06   0.956     .1927452    4.737273
------------------------------------------------------------------------------
```

```
. /// Descriptive statistics for censoring distribution

. bysort death: tabstat obstime, stat(n max) col(stat) by(ldlCTG)

-> death = 0
Summary for variables: obstime
```

```
     by categories of: ldlCTG

  ldlCTG |          N       max
---------+--------------------
       0 |         12  5.754962
      70 |        115  5.878166
     100 |        184  5.883641
     130 |        191  5.908282
     160 |         72  5.905544
     190 |         20  5.908282
---------+--------------------
   Total |        594  5.908282

-> death = 1
Summary for variables: obstime
     by categories of: ldlCTG

  ldlCTG |          N       max
---------+--------------------
       0 |         10  5.166325
      70 |         28  5.494866
     100 |         44  5.357974
     130 |         34  5.535934
     160 |         11   5.38809
     190 |          4  4.944559
---------+--------------------
   Total |        131  5.535934

. g censor = 1 - death

. stset obstime censor

     failure event:  censor != 0 & censor < .
obs. time interval:  (0, obstime]
 exit on or before:  failure

      735  total obs.
        0  exclusions
      735  obs. remaining, representing
      602  failures in single record/single failure data
 3630.376  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
```

```
                                    last observed exit t =   5.91102
```

```
. sts graph, t1("Censoring Distribution")

        failure _d:  censor
  analysis time _t:  obstime
```



```
. stci, p(10)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         10%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.029432      .0047079      5.01848    5.04038

. stci, p(25)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         25%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.086927      .0069813      5.07598    5.09788

. stci, p(50)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         50%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.185489      .0100226      5.16632    5.20465

. stci, p(75)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         75%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.664613      .0152841      5.62902    5.68652
```

```
. stci, p(90)

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of
             |  subjects         90%     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.776865      .0106913       5.7577    5.80424

. stci, rmean

        failure _d:  censor
  analysis time _t:  obstime

             |    no. of  restricted
             |  subjects        mean     Std. Err.     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       total |       735    5.332998      .0120774      5.30933    5.35667

. /// Descriptive statistics for LDL

. tabstat ldl, stat(n mean sd min q max) col(stat)

    variable |         N      mean        sd       min       p25       p50       p75       max
-------------+--------------------------------------------------------------------------------
         ldl |       725  125.8028  33.60197        11       102       125       147       247

. list obstime death if ldl==.

     |  obstime   death |
-----+------------------+
 17. |  5.91102       0 |
 34. | 5.054072       0 |
 87. | 5.670089       0 |
236. | 5.546885       0 |
511. |   5.7577       0 |
529. | 5.790555       0 |
555. | 5.262149       0 |
589. |  5.14716       0 |
608. | .6570842       1 |
700. | .1889117       1 |
```

```
. stcox i.ldlCTG, robust

         failure _d:  death
   analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(5)    =     15.42
Log pseudolikelihood =   -834.73192                 Prob > chi2     =    0.0087

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ldlCTG |
          70 |   .3980394   .1371452    -2.67   0.008     .2026017    .7820043
         100 |   .3925934   .1280949    -2.87   0.004     .2071162    .7441697
         130 |   .2939145   .0987499    -3.64   0.000     .1521363    .5678181
         160 |   .2565136   .1067636    -3.27   0.001      .113457    .5799484
         190 |    .316718   .1840332    -1.98   0.048     .1014077    .9891787
------------------------------------------------------------------------------

. predict dummyfit
(option hr assumed; relative hazard)
(10 missing values generated)

. egen grbg=mean(dummyfit) if ldl==160
(731 missing values generated)

. egen grbg2= mean(grbg)

. replace dummyfit= dummyfit / grbg2
(725 real changes made)
```

```
. drop grbg grbg2

. stcox ldl i.ldlCTG, robust

        failure _d:  death
  analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(6)    =     17.09
Log pseudolikelihood =   -834.43132                 Prob > chi2     =    0.0089

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   .9926048   .0089601    -0.82   0.411     .9751978    1.010322
             |
      ldlCTG |
          70 |   .5029001   .2374872    -1.46   0.146     .1993031    1.268964
         100 |   .6084833   .3904284    -0.77   0.439     .1730143    2.140007
         130 |   .5602902   .4835224    -0.67   0.502     .1032377    3.040799
         160 |   .6106519   .7247052    -0.42   0.678       .05965    6.251393
         190 |   .9738211   1.433135    -0.02   0.986      .054425    17.42449
------------------------------------------------------------------------------

. testparm i.ldl*

 ( 1)  70.ldlCTG = 0
 ( 2)  100.ldlCTG = 0
 ( 3)  130.ldlCTG = 0
 ( 4)  160.ldlCTG = 0
 ( 5)  190.ldlCTG = 0

           chi2(  5) =    5.14
```

```
         Prob > chi2 =    0.3988

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

. stcox ldl0 ldl70 ldl100 ldl130 ldl160 ldl190, robust

         failure _d:  death
   analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(6)    =     31.77
Log pseudolikelihood =   -833.80031                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ldl0 |   .9781009   .0092141    -2.35   0.019     .9602073    .9963279
       ldl70 |   .9797285   .0135753    -1.48   0.139     .9534795      1.0067
      ldl100 |   .9977076   .0109918    -0.21   0.835      .976395    1.019485
      ldl130 |    1.00361   .0125169     0.29   0.773      .979375    1.028445
      ldl160 |    .970916   .0214266    -1.34   0.181      .929816    1.013833
      ldl190 |   1.028804   .0259756     1.12   0.261     .9791323    1.080996
------------------------------------------------------------------------------

. predict splinefit
(option hr assumed; relative hazard)
(10 missing values generated)

. egen grbg=mean(splinefit) if ldl==160
(731 missing values generated)

. egen grbg2= mean(grbg)
```

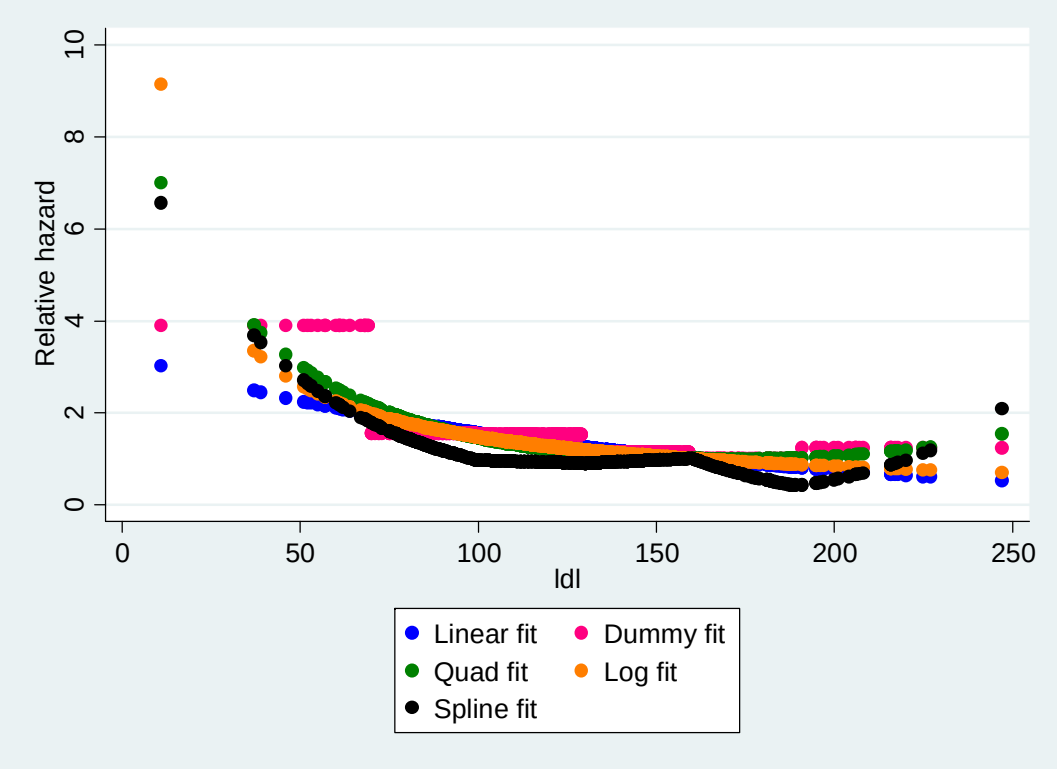
```
. replace splinefit= splinefit / grbg2
(725 real changes made)

. drop grbg grbg2

.
. test ldl0=ldl70=ldl100=ldl130=ldl160=ldl190

 ( 1)  ldl0 - ldl70 = 0
 ( 2)  ldl0 - ldl100 = 0
 ( 3)  ldl0 - ldl130 = 0
 ( 4)  ldl0 - ldl160 = 0
 ( 5)  ldl0 - ldl190 = 0

           chi2(  5) =    9.88
         Prob > chi2 =    0.0788

.
. stcox ldl ldl0 ldl70 ldl100 ldl130 ldl160 ldl190, robust

         failure _d:  death
   analysis time _t:  obstime

note: ldl190 omitted because of collinearity
Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =  3585.390827
                                                    Wald chi2(6)    =     31.77
Log pseudolikelihood =   -833.80031                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   1.028804   .0259756     1.12   0.261     .9791323    1.080996
        ldl0 |   .9507161   .0255971    -1.88   0.061     .9018476    1.002233
```

```
       ldl70 |   .9522981   .0273284    -1.70   0.089     .9002139    1.007396
      ldl100 |   .9697739   .0270032    -1.10   0.270     .9182668     1.02417
      ldl130 |   .9755113   .0261824    -0.92   0.356     .9255212    1.028201
      ldl160 |   .9437324   .0393203    -1.39   0.165     .8697289    1.024033
      ldl190 |  (omitted)
------------------------------------------------------------------------------

. test ldl0 ldl70 ldl100 ldl130 ldl160 ldl190

 ( 1)  ldl0 = 0
 ( 2)  ldl70 = 0
 ( 3)  ldl100 = 0
 ( 4)  ldl130 = 0
 ( 5)  ldl160 = 0
 ( 6)  o.ldl190 = 0
       Constraint 6 dropped

           chi2(  5) =    9.88
         Prob > chi2 =    0.0788

.

. stcox ldl

         failure _d:  death
   analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects =          725                     Number of obs   =       725
No. of failures =          131
Time at risk    =  3585.390827
                                                   LR chi2(1)      =      7.70
Log likelihood  =   -836.48275                     Prob > chi2     =    0.0055

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
         ldl |   .9926246   .0026735    -2.75   0.006     .9873984    .9978785
------------------------------------------------------------------------------

. predict linearfit
(option hr assumed; relative hazard)
(10 missing values generated)

. egen grbg=mean(linearfit) if ldl==160
(731 missing values generated)

. egen grbg2= mean(grbg)

. replace linearfit= linearfit / grbg2
(725 real changes made)

. drop grbg grbg2

. g logldl= log(ldl)
(10 missing values generated)

. stcox logldl

        failure _d:  death
  analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects =          725                     Number of obs   =       725
No. of failures =          131
Time at risk    =  3585.390827
                                                   LR chi2(1)      =      9.87
Log likelihood  =   -835.39584                     Prob > chi2     =    0.0017

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logldl |    .437521   .1046469    -3.46   0.001     .2737835    .6991824
------------------------------------------------------------------------------
```

```
. predict logfit
(option hr assumed; relative hazard)
(10 missing values generated)

. egen grbg=mean(logfit) if ldl==160
(731 missing values generated)

. egen grbg2= mean(grbg)

. replace logfit= logfit / grbg2
(725 real changes made)

. drop grbg grbg2

. g ldlsqr= ldl^2
(10 missing values generated)

. stcox ldl ldlsqr

        failure _d:  death
  analysis time _t:  obstime

Cox regression -- Breslow method for ties

No. of subjects =          725                     Number of obs   =       725
No. of failures =          131
Time at risk    =  3585.390827
                                                   LR chi2(2)      =     10.19
Log likelihood  =   -835.23455                     Prob > chi2     =    0.0061

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ldl |   .9742307   .0110009    -2.31   0.021     .9529062    .9960323
      ldlsqr |   1.000076   .0000449     1.70   0.089     .9999884    1.000164
------------------------------------------------------------------------------
```

```
. predict quadfit
(option hr assumed; relative hazard)
(10 missing values generated)

. egen grbg=mean(quadfit) if ldl==160
(731 missing values generated)

. egen grbg2= mean(grbg)

. replace quadfit= quadfit / grbg2
(725 real changes made)

. drop grbg grbg2

. twoway (scatter linearfit dummyfit quadfit logfit splinefit ldl, ///
>         color(blue pink green orange black yellow) ///
>         legend(label(1 "Linear fit") label(2 "Dummy fit") label(3 "Quad fit") ///
>         label(4 "Log fit") label(5 "Spline fit")))
```


Relative hazard
0
2
4
6
8
10
0
50
100
150
200
250
ldl
Linear fit
Dummy fit
Quad fit
Log fit
Spline fit