**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.

**In the logistic regression model used to answer this question, four distinct groups (Whites, Blacks, Asian, and other) are modeled with three regression parameters (three predictors and the intercept). It is a saturated model.**

* ***Methods:* The odds of subjects with a diabetes diagnosis were compared between four distinct race groups (Whites, Blacks, Asians, and Other), with Whites as a reference group, using a logistic regression model. Statistical inference on the odds of having diabetes was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates.**
* ***Results:* Data was available on 735 subjects, among which 572 subjects were Whites, 104 Blacks, 47 Asians, and 12 Others. There were 79 subjects (10.75%) with diabetes diagnosis, while 656 subjects not. From a logistic regression analysis, we estimate that the odds of having a diabetes diagnosis for White was 0.106. Based on a 95% confidence interval, this observed odds would not be judged unusual if a subject with the race as white had odds of anywhere from 0.082 to 0.143.** **A two-sided p value < 0.0001 suggests that we can with high confidence reject the null hypothesis that the odds of having a diabetes diagnosis for Whites is 0.**

**The odds ratio of having a diabetes diagnosis for Blacks compared to Whites was 1.929. Based on a 95% confidence interval, this observed odds ratio would not be judged unusual if a subject with the race as black compared to Whites had odds ratio of anywhere from 1.082 to 3.439.** **A two-sided p value 0.026 suggests that we can with high confidence reject the null hypothesis that Blacks has the same odds of having a diabetes diagnosis with Whites.**

**The odds ratio of having a diabetes diagnosis for Asians compared to Whites was 0.628. Based on a 95% confidence interval, this observed odds ratio would not be judged unusual if a subject with the race as Asians compared to Whites had odds ratio of anywhere from 0.189 to 2.090.** **A two-sided p value 0.449 suggests that we cannot with high confidence reject the null hypothesis that Asians has the same odds of having a diabetes diagnosis with Whites.**

**The odds ratio of having a diabetes diagnosis for other racial group compared to Whites was 1.843. Based on a 95% confidence interval, this observed odds ratio would not be judged unusual if a subject with the race as other compared to Whites had odds ratio of anywhere from 0.393 to 8.631.** **A two-sided p value 0.438 suggests that we cannot with high confidence reject the null hypothesis that the other racial group has the same odds of having a diabetes diagnosis with Whites.**

**The two-sided p value for chi-square test is 0.0956 suggests that we cannot with high confidence reject the null hypothesis that race is not associated with odds of having a diabetes diagnosis.**

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

**The intercept corresponds to the odds of having diabetes diagnosis for Whites. The estimated odds is 0.106. (95% CI 0.082-0.143). The two-sided p value < 0.0001 suggests the odds is highly statistically different from 0.**

**The slope for “black” corresponds to the odds ratio of having diabetes diagnosis for Blacks compared to Whites. The estimated odds ratio is 1.929. (95% CI 1.082 - 3.439). The two-sided p value 0.026 suggests the odds ratio is highly statistically different from 1.**

**The slope for “Asian” corresponds to the odds ratio of having diabetes diagnosis for Asians compared to Whites. The estimated odds ratio is 0.628. (95% CI 0.189 - 2.090). The two-sided p value 0.449 suggests the odds ratio is not highly statistically different from 1.**

**The slope for “other” corresponds to the odds ratio of having diabetes diagnosis for Asians compared to Whites. The estimated odds ratio is 1.843. (95% CI 0.393 - 8.631). The two-sided p value 0.438 suggests the odds ratio is not highly statistically different from 1.**

* 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.

**The two-sided p value from the intercept < 0.0001 suggests the odds for whites is highly statistically different from 0.**

**The two-sided p value from the slope for black 0.026 also suggests the odds ratio for blacks compared to whites is highly statistically different from 1.**

**The two-sided p values from the slope for Asians and other are 0.449 and 0.438 respectively, which suggest the odds ratios for Asians and Other compared to Whites are not highly statistically different from 1.**

**The two-sided p value for chi-square test is 0.0956 suggests that the association between race and odds of having diabetes diagnosis is not highly statistically significant.**

* 1. Now fit a logistic regression model that uses blacks as a reference group. How would your report of formal inference differ from that that you provided in part (a)? How does this regression model relate to that in part (a)?
* ***Methods:* The odds of subjects with a diabetes diagnosis were compared between four distinct race groups (Whites, Blacks, Asians, and Other), with Blacks as a reference group, using a logistic regression model. Statistical inference on the odds of having diabetes was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates.**
* ***Results:* Data was available on 735 subjects, among which 572 subjects were Whites, 104 Blacks, 47 Asians, and 12 Others. There were 79 subjects (10.75%) with diabetes diagnosis, while 656 subjects not. From a logistic regression analysis, we estimate that the odds of having a diabetes diagnosis for Blacks was 0.209. Based on a 95% confidence interval, this observed odds would not be judged unusual if a subject with the race as Blacks had odds of anywhere from 0.126 to 0.348.** **A two-sided p value < 0.0001 suggests that we can with high confidence reject the null hypothesis that the odds of having a diabetes diagnosis for Blacks is 0.**

**The odds ratio of having a diabetes diagnosis for Whites compared to Blacks was 0.519. Based on a 95% confidence interval, this observed odds ratio would not be judged unusual if a subject with the race as white compared to Blacks had odds ratio of anywhere from 0.291 to 0.925. A two-sided p value 0.026 suggests that we can with high confidence reject the null hypothesis that Whites has the same odds of having a diabetes diagnosis with Blacks.**

**The odds ratio of having a diabetes diagnosis for Asians compared to Blacks was 0.326. Based on a 95% confidence interval, this observed odds ratio would not be judged unusual if a subject with the race as Asians compared to Blacks had odds ratio of anywhere from 0.091 to 1.167. A two-sided p value 0.085 suggests that we cannot with high confidence reject the null hypothesis that Asians has the same odds of having a diabetes diagnosis with Blacks.**

**The odds ratio of having a diabetes diagnosis for other racial group compared to Blacks was 0.956. Based on a 95% confidence interval, this observed odds ratio would not be judged unusual if a subject with the race as other compared to Blacks had odds ratio of anywhere from 0.193 to 4.742. A two-sided p value 0.956 suggests that we cannot with high confidence reject the null hypothesis that the other racial group has the same odds of having a diabetes diagnosis with Blacks.**

**The two-sided p value for chi-square test is 0.0956 suggests that we cannot with high confidence reject the null hypothesis that race is not associated with odds of having a diabetes diagnosis.**

**Although both models are saturate models, the reference groups for two models are different. Because the reference group was changed, the meaning of intercept and slopes were changed. The intercepts were the estimated odds for the reference groups (the first one: Whites and the second one: Blacks), and the slopes were the estimated odds ratios for each racial groups compared to the reference groups. Consequently, the 95% CIs and p values were changed according to their estimates and standard error. However, the two-sided p values for chi-square test were the same between two models.**

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

**The intercept corresponds to the odds of having diabetes diagnosis for Blacks. The estimated odds is 0.209. (95% CI 0.082-0.348). The two-sided p value < 0.0001 suggests the odds is highly statistically different from 0.**

**The slope for “white” corresponds to the odds ratio of having diabetes diagnosis for Whites compared to Blacks. The estimated odds ratio is 0.519. (95% CI 0.291 - 0.925). The two-sided p value 0.026 suggests the odds ratio is highly statistically different from 1.**

**The slope for “Asian” corresponds to the odds ratio of having diabetes diagnosis for Asians compared to Blacks. The estimated odds ratio is 0.326. (95% CI 0.091 - 1.167). The two-sided p value 0.085 suggests the odds ratio is not highly statistically different from 1.**

**The slope for “other” corresponds to the odds ratio of having diabetes diagnosis for Asians compared to Blacks. The estimated odds ratio is 0.956. (95% CI 0.193 - 4.742). The two-sided p value 0.956 suggests the odds ratio is not highly statistically different from 1.**

* 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.

**The two-sided p value from the intercept < 0.0001 suggests the odds for blacks is highly statistically different from 0.**

**The two-sided p value from the slope for black 0.026 also suggests the odds ratio for Whites compared to Blacks is highly statistically different from 1.**

**The two-sided p values from the slope for Asians and other are 0.085 and 0.956 respectively, which suggest the odds ratios for Asians and Other compared to Blacks are not highly statistically different from 1.**

**The two-sided p value for chi-square test is 0.0956 suggests that the association between race and odds of having diabetes diagnosis is not highly statistically significant.**

* 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)?

**The p-values for intercept and each slope are the significant test for the odds for reference groups and the odds ratio for each racial group compared to the reference groups, respectively. As reference groups change, the p value for each parameter will be changed. Even the p value for one regression parameter is not significantly, it doesn’t mean that the dummy variable (as a reference group or the group compared to the reference group) has statistical non-significance in the regression model.**

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:* Serum LDL was categorized into 6 groups 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. Distributions of time to death from any cause was estimated using Kaplan-Meier estimates with strata defined by the above categorized serum LDL. Quantification of association between all-cause mortality was summarized by the hazards ratio computed from the proportional regression model (using LDL less than 70g/dL as reference group), with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.**
* ***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. There were 22 subjects in the group with LDL less than 70 mg/dL, 143 in the group with LDL 70-99 mg/dL, 228 in the group with LDL 100-129 mg/dL, 225 in the group with LDL 130-159 mg/dL, 82 in the group with LDL 160-189 mg/dL, and 24 in the group with LDL greater than or equal to 190 mg/dL.**

**From a proportional hazards regression analysis, we estimate that the instantaneous risk of death for the group with serum LDL 70-99 mg/dL is a relative 60.20% lower (hazard ratio 0.398) compared to the group with serum LDL less than 70mg/dL. Based on a 95% confidence interval, this observed hazard ratio of 0.398 for the comparison of the group with LDL 70-99 mg/dL to the group with LDL less than 70mg/dL would not be judged unusual if the true hazard ratio were anywhere between 0.203 to 0.782. A two-sided p value P 0.008 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is the same for the higher and lower LDL groups.**

**The instantaneous risk of death for the group with serum LDL 100-129 mg/dL is a relative 60.74% lower (hazard ratio 0.393) compared to the group with serum LDL less than 70mg/dL. Based on a 95% confidence interval, this observed hazard ratio of 0.393 for the comparison of the group with LDL 100-129 mg/dL to the group with LDL less than 70mg/dL would not be judged unusual if the true hazard ratio were anywhere between 0.207 to 0.744. A two-sided p value P 0.004 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is the same for the higher and lower LDL groups.**

**The instantaneous risk of death for the group with serum LDL 130-159 mg/dL is a relative 70.61% lower (hazard ratio 0.294) compared to the group with serum LDL less than 70mg/dL. Based on a 95% confidence interval, this observed hazard ratio of 0.294 for the comparison of the group with LDL 130-159 mg/dL to the group with LDL less than 70mg/dL would not be judged unusual if the true hazard ratio were anywhere between 0.152 to 0.568. A two-sided p value P <0.0001 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is the same for the higher and lower LDL groups.**

**The instantaneous risk of death for the group with serum LDL 160-189 mg/dL is a relative 74.35% lower (hazard ratio 0.257) compared to the group with serum LDL less than 70mg/dL. Based on a 95% confidence interval, this observed hazard ratio of 0.257 for the comparison of the group with LDL 160-189 mg/dL to the group with LDL less than 70mg/dL would not be judged unusual if the true hazard ratio were anywhere between 0.113 to 0.580. A two-sided p value P 0.001 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is the same for the higher and lower LDL groups.**

**The instantaneous risk of death for the group with serum LDL ≥190 mg/dL is a relative 68.33% lower (hazard ratio 0.317) compared to the group with serum LDL less than 70mg/dL. Based on a 95% confidence interval, this observed hazard ratio of 0.317 for the comparison of the group with LDL ≥190 mg/dL to the group with LDL less than 70mg/dL would not be judged unusual if the true hazard ratio were anywhere between 0.101 to 0.989. A two-sided p value P 0.048 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is the same for the higher and lower LDL groups.**

**The two-sided p value for chi-square test is 0.0087 suggests that we can with high confidence reject the null hypothesis that there is a straight line association across groups.**

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

**The intercept corresponds to the hazard of death for the group with serum LDL less than 70 mg/dL, which didn’t showed in the stata report.**

**The slope for “70” corresponds to the hazard ratio of death (0.398) for the comparison of the group with LDL 70-99 mg/dL to the group with LDL less than 70mg/dL. (95% CI 0.203 - 0.782). The two-sided p value 0.008 suggests the hazard ratio is highly statistically different from 1.**

**The slope for “100” corresponds to the hazard ratio of death (0.393) for the comparison of the group with LDL 100-129 mg/dL to the group with LDL less than 70mg/dL. (95% CI 0.207 - 0.744). The two-sided p value 0.004 suggests the hazard ratio is highly statistically different from 1.**

**The slope for “130” corresponds to the hazard ratio of death (0.294) for the comparison of the group with LDL 130-159 mg/dL to the group with LDL less than 70mg/dL. (95% CI 0.152 - 0.568). The two-sided p value <0.0001 suggests the hazard ratio is highly statistically different from 1.**

**The slope for “160” corresponds to the hazard ratio of death (0.257) for the comparison of the group with LDL 160-189 mg/dL to the group with LDL less than 70mg/dL. (95% CI 0.113 - 0.580). The two-sided p value 0.001 suggests the hazard ratio is highly statistically different from 1.**

**The slope for “190” corresponds to the hazard ratio of death (0.317) for the comparison of the group with LDL ≥190 mg/dL to the group with LDL less than 70mg/dL. (95% CI 0.101 - 0.989). The two-sided p value 0.048 suggests the hazard ratio is highly statistically different from 1.**

**The two-sided p value for chi-square test is 0.0087 suggests that we can with high confidence reject the null hypothesis that there is a straight line association across groups.**

* 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?

**Using testparm command in Stata to perform F test of linear restriction applied to the most recently fit model. It performed five test to assume the parameter coefficient of each five LDL group equal to zero (except the reference LDL group (<70 mg/dL)). The two-sided p value is 0.0087, which suggests that there is a nonlinearity across groups.**

* 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:* Serum LDL was categorized into 6 groups 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. Distributions of time to death from any cause was estimated using Kaplan-Meier estimates within each strata defined by the above categorized serum LDL. Quantification of association between all-cause mortality was summarized by the hazards ratio computed from the proportional hazard regression modeling serum LDL as a continuous random variable within each stratum, with the lowest LDL level within each stratum as the reference point. Confidence intervals and two-sided p values were computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.**
* ***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. There were 22 subjects in the group with LDL less than 70 mg/dL, 143 in the group with LDL 70-99 mg/dL, 228 in the group with LDL 100-129 mg/dL, 225 in the group with LDL 130-159 mg/dL, 82 in the group with LDL 160-189 mg/dL, and 24 in the group with LDL greater than or equal to 190 mg/dL.**

**From a proportional hazards regression analysis, we estimate that within the strata from LDL 0 to 69 mg/dL, the instantaneous risk of death is a relative 2.19% lower (hazard ratio 0.978) for each 1 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio suggesting lower death rates for groups of patients with higher LDL levels would not be judged unusual if the true instantaneous risk of death were anywhere from 0.37% to 3.98% lower in a group having serum LDL 1 mg/dL higher than that in another group (95% CI for hazard ratio 0.960 to 0.996). A two-sided p value of 0.019 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.**

**Within the strata from LDL 70 to 99 mg/dL, the instantaneous risk of death is a relative 2.03% lower (hazard ratio 0.980) for each 1 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio would not be judged unusual if the true instantaneous risk of death were anywhere from 0.67% higher to 4.65% lower in a group having serum LDL 1 mg/dL higher than that in another group (95% CI for hazard ratio 0.953 to 1.007). A two-sided p value of 0.139 suggests that we cannot with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL level.**

**Within the strata from LDL 100 to 129 mg/dL, the instantaneous risk of death is a relative 0.23% lower (hazard ratio 0.998) for each 1 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio would not be judged unusual if the true instantaneous risk of death were anywhere from 1.95% higher to 2.36% lower in a group having serum LDL 1 mg/dL higher than that in another group (95% CI for hazard ratio 0.976 to 1.019). A two-sided p value of 0.835 suggests that we cannot with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL level.**

**Within the strata from LDL 130 to 159 mg/dL, the instantaneous risk of death is a relative 0.36% higher (hazard ratio 1.004) for each 1 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio would not be judged unusual if the true instantaneous risk of death were anywhere from 2.84% higher to 2.06% lower in a group having serum LDL 1 mg/dL higher than that in another group (95% CI for hazard ratio 0.979 to 1.028). A two-sided p value of 0.773 suggests that we cannot with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL level.**

**Within the strata from LDL 160 to 189 mg/dL, the instantaneous risk of death is a relative 2.91% lower (hazard ratio 0.971) for each 1 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio would not be judged unusual if the true instantaneous risk of death were anywhere from 1.38% higher to 7.02% lower in a group having serum LDL 1 mg/dL higher than that in another group (95% CI for hazard ratio 0.930 to 1.014). A two-sided p value of 0.181 suggests that we cannot with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL level.**

**Within the strata with LDL ≥190 mg/dL, the instantaneous risk of death is a relative 2.88% higher (hazard ratio 1.029) for each 1 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio would not be judged unusual if the true instantaneous risk of death were anywhere from 8.10% higher to 2.09% lower in a group having serum LDL 1 mg/dL higher than that in another group (95% CI for hazard ratio 0.979 to 1.01). A two-sided p value of 0.261 suggests that we cannot with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL level.**

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

**The intercept corresponds to the hazard of death for the group with serum LDL equal to 0 mg/dL, which didn’t showed in the stata report.**

**The slope for “ldl0” corresponds to the hazard ratio of death (0. 978) for the group having serum LDL 1 mg/dL higher compared to that in another group within the range of LDL between 0-69 mg/dL. (95% CI 0.960 to 0.996). The two-sided p value 0.019 suggests the hazard ratio is highly statistically different from 1.**

**The slope for “ldl70” corresponds to the hazard ratio of death (0. 980) for the group having serum LDL 1 mg/dL higher compared to that in another group within the range of LDL between 70-99 mg/dL. (95% CI 0.953 to 1.007). The two-sided p value 0.139 suggests the hazard ratio is not highly statistically different from 1.**

**The slope for “ldl100” corresponds to the hazard ratio of death (0.998) for the group having serum LDL 1 mg/dL higher compared to that in another group within the range of LDL between 100-129 mg/dL. (95% CI 0.976 to 1.019). The two-sided p value 0.835 suggests the hazard ratio is not highly statistically different from 1.**

**The slope for “ldl130” corresponds to the hazard ratio of death (1.004) for the group having serum LDL 1 mg/dL higher compared to that in another group within the range of LDL between 130-159 mg/dL. (95% CI 0.979 to 1.028). The two-sided p value 0.773 suggests the hazard ratio is not highly statistically different from 1.**

**The slope for “ldl160” corresponds to the hazard ratio of death (0.971) for the group having serum LDL 1 mg/dL higher compared to that in another group within the range of LDL between 160-189 mg/dL. (95% CI 0.930 to 1.014). The two-sided p value 0.181 suggests the hazard ratio is not highly statistically different from 1.**

**The slope for “ldl190” corresponds to the hazard ratio of death (1.029) for the group having serum LDL 1 mg/dL higher compared to that in another group within the range of LDL ≥190 mg/dL. (95% CI 0.979 to 1.01). The two-sided p value 0.261 suggests the hazard ratio is not highly statistically different from 1.**

* 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?

**Using test command in Stata to perform F test of linear restriction applied to the most recently fit model. It performed five test to compared five LDL groups to the ldl0 group. The null hypotheses was that all the parameter coefficients would have to be equal. The two-sided p value is 0.0788. There is no statistically significant evidence to prove that there is a nonlinearity across groups.**

* 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 used in homeworks 4 and 5 were survival analysis, which employed all of the information the data could provide (those who died within 5 years and survived after 5 years). Furthermore, the analyses homeworks 4 and 5 were conditioning on the variable we might think of as a “cause” (serum LDL) and consider the distribution of the putative “effect” (hazard of death over the entire period of observation). This seems scientifically more pleasing.**

* 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.

**We can see from the below figure that the model fit from problem 2 has five different values of relative hazard ratio compared to that of ldl 160. Instead, the model fit from problem 3 has relatively smooth curve jointed by some knots.**

* 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?

**I would prefer to use proportional hazard regression modeling serum LDL as a continuous untransformed random variable. Because it is not statistically significant to prove that there is not a nonlinearity across different LDL values. In this situation, it is better not to dichotomize the predictor of interest.**

**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.