**Homework #5**

February 3, 2014

1. Perform a statistical regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across race.
	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: Descriptive statistics were generated to summarize the data according to variables of clinical interest in diabetes, including diabetes diagnosis at baseline, age, weight, height, sex, and physical activity reported for the week prior to study entry, measured in 1,000 kcal. The mean, standard deviation (SD), minimum, and maximum were calculated according to race (white, black, Asian, or other), and are presented in Table 1. Observation time is a censored variable and descriptive statistics were determined by Kaplan-Meier estimates for the median survival time for each group according to race. For inferential statistical analysis, the odds of baseline diabetes diagnoses were compared across groups defined by race by using logistic regression with robust standard error estimation (Huber-White sandwich estimator). Odds were determined by maximum likelihood estimation from the regression analysis, with Wald-based 95% confidence intervals (CI) and two-sided p-values reported for each parameter estimate. The variable for diabetes diagnosis is binary, while the variable for race is categorical. Race was coded as a dummy variable such that white, black, Asian, and other were indicator variables according to race, and “white” was used as the reference group. Statistical significance is defined by the threshold of 0.05. To test for nonlinearity, a Wald test was conducted on the races groups modeled in the regression analysis. This model is saturated as there are four groups (race) and four independent parameters (slope for black, Asian, and other races comparing odds ratio to that of white race, and intercept).

Table 1. Demographic information for subjects according to self-reported race category

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **White****(n = 572)** | **Black****(n = 104)** | **Asian****(n = 47)** | **Other****(n = 12)** | **Any Race****(n = 735)** |
| Diabetes, prop. (SD) | 0.098 (0.297) | 0.173 (0.380) | 0.064 (0.247) | 0.167 (0.389) | 0.107 (0.310) |
| Mean Age, years (SD, range) | 74.4 (5.44, 65-99) | 75.1(5.51, 65-90) | 74.9(5.20, 67-89) | 74.9(6.68, 68-91) | 74.6(5.45, 65-99) |
| Male, % (SD) | 0.500 (0.5004) | 0.490 (0.502) | 0.447 (0.503) | 0.667 (0.492) | 0.498 (0.5003) |
| Mean Weight, lb (SD, range) | 159.3 (29.9, 74.0-253.0) | 162.1(35.5, 96.0-258.0) | 160.4 (31.4, 96.0-264.0) | 169.3(26.4, 137.0-222.0) | 159.9(30.7, 74.0-264.0) |
| Mean Height, cm (SD, range) | 165.8(9.61, 141-191) | 165.4 (10.8, 139-188) | 166.4(9.44, 150-187) | 166.4(5.03, 158-175) | 165.8(9.71, 139-191) |
| Physical activity, mean kcal/week, (SD, range) | 1956 (2126, 0-13,815) | 1700(1710, 0-9840) | 2180(2002, 0-7410) | 1237(930, 0-3030) | 1922(2052, 0-13,815) |
| Median survival, years (95% CI) | 5.19(5.17, 5.22) | 5.17(5.12, 5.20) | 5.18(5.15, 5.53) | 5.16(5.03, 5.75) | 5.19(5.17, 5.20) |

Abbreviations: cm = centimeters, kcal = kilocalories lb = pounds, SD = standard deviation

Inference: Data regarding race and diabetes status were available for all 735 subjects in this sample. Demographic characteristics of clinical interest to diabetes were similar for all race groups, except the proportions of subjects in the “white” and “Asian” race groups were lower than those in the “black” or “other” race groups. The mean observation time of the study was similar between groups, except in the “other” race category the mean observation time was shorter with a larger standard deviation due to the small sample size in that group.

 By logistic regression the log odds for a diabetes diagnosis in the white group is -2.22 (95% CI -2.50, -1.94). Taking the antilogarithmic transformation e^-2.22, the odds ratio of diabetes in the white group is 0.109, which is consistent with a true population odds ratio from 0.082 to 0.143. In the black group at baseline the odds ratio was 1.93 (95% CI 1.082, 3.439), that is, the odds of diabetes is 92.9% higher among subjects in the black group compared to the white group, which would not be surprising if the true population odds was between 8.15% and 344% higher in the black group. The odds ratio for a diabetes diagnosis in the Asian group at baseline was 0.628 (95% CI 0.189, 2.09), that is, the odds of diabetes is 37.2% lower among subjects in the Asian group compared to the white group, which would not be surprising if the true population odds was between 81.1% lower and 209% higher in the Asian group. The odds ratio for a diabetes diagnosis in the “other” race group at baseline was 1.84 (95% CI 0.393, 8.63), that is, the odds of diabetes is 84.3% higher among subjects in the “other” group compared to the white group, which would not be surprising if the true population odds was between 60.7% lower and 863% higher in the “other” group. The association between race and diabetes diagnosis at study baseline is not statistically significant (two-sided p=0.0956). This suggests we have insufficient evidence to reject the null hypothesis that the odds of diabetes is the same across race groups.

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

There are four parameters in total:

* Slope for “black” race: This parameter is the difference in the log odds of a diabetes diagnosis in the black group compared to that of the white group, since white is the reference group. According to this analysis, the log odds of diabetes in the black group is 0.657 larger than the log odds of diabetes in the white group (95% CI 0.078, 1.24).
* Slope for “Asian” race: This parameter is the difference in the log odds of a diabetes diagnosis in the Asian group compared to that of the white group, since white is the reference group. According to this analysis, the log odds of diabetes in the Asian group is 0.465 smaller than the log odds of diabetes in the white group (95% CI -1.67, 0.738).
* Slope for “other” race: This parameter is the difference in the log odds of a diabetes diagnosis in the “other” group compared to that of the white group, since white is the reference group. According to this analysis, the log odds of diabetes in the “other” group is 0.611 larger than the log odds of diabetes in the white group (95% CI -0.933, 2.16).
* Intercept: This parameter corresponds to the log odds for the white group. According to the analysis, the log odds of diabetes in the white group is -2.22 (95% CI -2.50, -1.94).
	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.

Slope for “black” race: According to this analysis, the odds of diabetes in the black group is 1.93 times the odds of diabetes in the white group (95% CI 1.08, 3.43) with a p-value of 0.026. Ignoring multiple comparisons, this would suggest there is sufficient evidence to reject the null hypothesis of no association between diabetes and black race in favor of the alternative hypothesis that black race increases the odds of diabetes compared to the white race.

Slope for “Asian” race: According to this analysis, the odds of diabetes in the Asian group is 0.628 times the odds of diabetes in the white group (95% CI 0.189, 2.09) with a p-value of 0.449. This suggests insufficient evidence to reject the null hypothesis that there is no association between diabetes and Asian race relative to white race.

Slope for “other” race: According to this analysis, the odds of diabetes in the “other” group is 1.84 times the odds of diabetes in the white group (95% CI 0.393, 8.63), with a p-value of 0.438. This suggests insufficient evidence to reject the null hypothesis that there is no association between diabetes and “other” race relative to white race.

* 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)?

Compared to (a) the inference for this model would be compared to the odds of subjects in the black group being diagnosed with diabetes. The following differences would occur:

Methods: For inferential statistical analysis, the odds of baseline diabetes diagnoses were compared across groups defined by race by using logistic regression with robust standard error estimation (Huber-White sandwich estimator). Odds were determined by maximum likelihood estimation from the regression analysis, with Wald-based 95% confidence intervals (CI) and two-sided p-values reported for each parameter estimate. The variable for diabetes diagnosis is binary, while the variable for race is categorical. Race was coded as a dummy variable such that white, black, Asian, and other were indicator variables according to race, and **“black” was used as the reference group**. Statistical significance is defined by the threshold of 0.05. To test for nonlinearity, a Wald test was conducted on the races groups modeled in the regression analysis.

Inference: Data regarding race and diabetes status were available for all 735 subjects in this sample. By logistic regression the log odds for a diabetes diagnosis in the black group is -1.56 (95% CI -2.07, -1.06). Taking the antilogarithmic transformation e^-1.56, the odds ratio of diabetes in the black group is 0.209, which is consistent with a true population odds ratio from 0.126 to 0.348. In the white group at baseline, the odds ratio was 0.519 (95% CI 0.291, 0.925), that is, the odds of diabetes is 48.1% lower among subjects in the white group compared to the black reference group, which would not be surprising if the true population odds was between 7.5% and 70.9% lower in the white group. The odds ratio for a diabetes diagnosis in the Asian group at baseline was 0.326 (95% CI 0.909, 1.17), that is, the odds of diabetes is 67.4% lower among subjects in the Asian group compared to the black group, which would not be surprising if the true population odds was between 9.1% lower and 16.7% higher in the Asian group. The odds ratio for a diabetes diagnosis in the “other” race group at baseline was 0.956 (95% CI 0.193, 4.74), that is, the odds of diabetes is 4.4% lower among subjects in the “other” group compared to the black group, which would not be surprising if the true population odds was between 80.7% lower and 474% higher in the “other” group. The association between race and diabetes diagnosis at study baseline is not statistically significant (two-sided p=0.0956). This suggests we have insufficient evidence to reject the null hypothesis that the odds of diabetes is the same across race groups.

* 1. Using the regression model fit in part (d), provide an interpretation for each of the regression parameters (including the intercept.)
* Slope for “white” race: This parameter is the difference in the log odds of a diabetes diagnosis in the white group compared to that of the black reference group. According to this analysis, the log odds of diabetes in the white group is 0.657 smaller than the log odds of diabetes in the black group (95% CI -1.24, -0.078).
* Slope for “Asian” race: This parameter is the difference in the log odds of a diabetes diagnosis in the Asian group compared to that of the black reference group. According to this analysis, the log odds of diabetes in the Asian group is 1.12 smaller than the log odds of diabetes in the black group (95% CI -2.40, 0.154).
* Slope for “other” race: This parameter is the difference in the log odds of a diabetes diagnosis in the “other” group compared to that of the black reference group. According to this analysis, the log odds of diabetes in the “other” group is 0.045 smaller than the log odds of diabetes in the black group (95% CI -1.65, 1.56).
* Intercept: This parameter corresponds to the log odds for the black group. According to the analysis, the log odds of diabetes in the black group is -1.56 (95% CI -2.07, -1.05).
	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.

Slope for “white” race: According to this analysis, the odds of diabetes in the white group is 0.519 times the odds of diabetes in the black group (95% CI 0.291, 0.925) with a p-value of 0.026. Ignoring multiple comparisons, this would suggest there is sufficient evidence to reject the null hypothesis of no association between diabetes and white race in favor of the alternative hypothesis that black race increases the odds of diabetes compared to the white race.

Slope for “Asian” race: According to this analysis, the odds of diabetes in the Asian group is 0.326 times the odds of diabetes in the black group (95% CI 0.091, 1.17) with a p-value of 0.085. This suggests insufficient evidence to reject the null hypothesis that there is no association between diabetes and Asian race relative to black race.

Slope for “other” race: According to this analysis, the odds of diabetes in the “other” group is 0.956 times the odds of diabetes in the white group (95% CI 0.193, 4.74), with a p-value of 0.956. This suggests insufficient evidence to reject the null hypothesis that there is no association between diabetes and “other” race relative to black race.

* 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)?

This is a situation in which the p-values depend very much upon the reference case, and changing the reference case can change the outcome of the analysis. It is much more appropriate and robust to decide *a priori* what variables are of interest. Since dummy variables essentially fit a step-function to the data, precision is lost, and ordering of the variables is lost which eliminates the ability to detect trends in the data. Since it is a step function, it is likely to increase the root mean squared error (RMSE) estimate, reducing precision. If the data is likely to be well approximated by a linear fit, it is much more efficient to do the linear fit rather than use dummy variables.

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

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

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

Methods: Since the variable for observation time (time to death or study termination) is censored, descriptive statistics were generated using Kaplan-Meier techniques to summarize the data according to the median survival time, minimum, maximum, the probability of surviving two years, and the probability of surviving five years described for groups according to the Mayo Clinic guidelines for LDL risk (LDL < 70 mg/dL; 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and ≥190 mg/dL). The Kaplan-Meier survival curves were displayed on a graph according to LDL category. In addition, descriptive statistics to compare the sample characteristics in each LDL group were calculated, including mean LDL measurement within the grouped range, mean age, sex, and number of deaths.

For inferential statistical analysis, the instantaneous risk (hazard) of death during the observation period was compared across groups defined by LDL category by using proportional hazards regression with robust standard error estimation (Huber-White sandwich estimator). Parameter estimates were determined by maximum likelihood estimation from the regression analysis, with Wald-based 95% confidence intervals (CI) and two-sided p-values reported for each parameter estimate. Serum LDL measurements were categorized according to Mayo Clinic guidelines and were untransofrmed. Subjects with missing serum LDL measurements were omitted from this analysis. Statistical significance was defined by the threshold of 0.05.

Descriptive Statistics: Of a total of 735 study subjects, serum LDL measurements were available for 725. Of the 10 subjects with missing LDL measurements, two died during the study period, and the remaining 8 were still alive at study termination. Table 2 lists the descriptive statistics of interest for the subjects with complete LDL data in this study. Figure 1 illustrates the survival progression of subjects in each LDL category during study observation.

Table 2. Demographic information for subjects according to measured serum LDL category

|  |  |
| --- | --- |
|  | **Baseline Serum LDL (mg/dL)** |
|  | **11-69** **(n=22)** | **70-99** **(n=143)** | **100-129** **(n=228)** | **130-159** **(n=225)** | **160-189** **(n=83)** | **190-247** **(n=24)** | **All Subjects1** **(n=725)** |
| **Deaths, n** | 10 | 28 | 44 | 34 | 11 | 7 | 131 |
| **Median survival, years (95% CI)** | 5.23 (5.13, 5.57) | 5.16 (5.12, 5.57) | 5.18 (5.14, 5.23) | 5.20 (5.17, 5.26) | 5.16 (5.13, 5.23) | 5.58 (5.15. 5.74) | 5.19 (5.17, 5.20) |
| **Maximum observation time, years** | 5.75 | 5.88 | 5.88 | 5.91 | 5.91 | 5.91 | 5.91 |
| **2-year Survival, probability** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 95.4% |
| **5-year Survival, probability** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 83.5% |
| **Mean LDL, mg/dL (SD)** | 56.2 (13.6) | 86.7 (8.25) | 114.7 (8.36) | 142.7 (8.53) | 172.3 (9.21) | 208.3 (13.5) | 125.8 (33.6) |
| **Mean age, years (SD)** | 75.5 (5.63) | 74.7 (5.49) | 74.6 (5.08) | 74.2 (5.62) | 74.6 (5.67) | 76.0 (6.11) | 74.6 (5.45) |
| **Male** | 68.2% | 53.9% | 53.9% | 43.1% | 48.2% | 20.8% | 49.7% |

1Only subjects with available serum LDL measurements were included in the study



Figure 1. Kaplan-Meier survival curves for subjects with available serum LDL measurements, according to LDL category.

Inference: Of 735 subjects in this study, data regarding serum LDL measurements were available for 725 subjects. By proportional hazards regression the hazard for the group with LDL 70-99 mg/dL is 60.2% lower (hazard ratio 0.398) than those in the group with LDL <70 mg/dL, which would not be surprising if the true hazard was within the 95% CI from 21.8% to 79.7% lower. For the group with LDL 100-129 mg/dL, the hazard of death is 60.7% lower (hazard ratio 0.393) than those in the group with LDL <70 mg/dL, which would not be surprising if the true hazard was within the 95% CI from 25.6% to 79.3% lower. For the group with LDL 130-159 mg/dL, the hazard of death is 70.6% lower (hazard ratio 0.294) than those in the group with LDL <70 mg/dL, which would not be surprising if the true hazard was within the 95% CI from 43.2% to 84.8%% lower. For the group with LDL 160-189 mg/dL, the hazard of death is 73.2% lower (hazard ratio 0.257) than those in the group with LDL <70 mg/dL, which would not be surprising if the true hazard was within the 95% CI from 42.0% to 88.7% lower. For the group with LDL greater than 190 mg/dL, the hazard of death is 68.3% lower (hazard ratio 0.317) than those in the group with LDL <70 mg/dL, which would not be surprising if the true hazard was within the 95% CI from 1.08% to 89.9% lower. The association between serum LDL categories at baseline and 5-year all-cause mortality is statistically significant (two-sided p=0.0087). This suggests we have sufficient evidence to reject the null hypothesis that the hazard of death is the same across LDL groups. This also suggests significant evidence of nonlinearity in the data.

3/3 for descriptive statistics

3/3 for performing an appropriate analysis

4/4 for reporting the association appropriately

Total: 10

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.
* Slope for LDL 70: This parameter is the hazard ratio of death in the group with LDL = 70-99 mg/dL compared to that of the group with LDL <70 mg/dL. According to this analysis, the hazard ratio in the “LDL 70” group is 60.2% lower (hazard ratio 0.398) than that of the “LDL 0” group (95% CI hazard ratio 0.203, 0.782).
* Slope for LDL 100: This parameter is the hazard ratio of death in the group with LDL = 100-129 mg/dL compared to that of the group with LDL <70 mg/dL. According to this analysis, the hazard ratio in the “LDL 100” group is 60.3% lower (hazard ratio 0.393) than that of the “LDL 0” group (95% CI hazard ratio 0.207, 0.744).
* Slope for LDL 130: This parameter is the hazard ratio of death in the group with LDL = 130-159 mg/dL compared to that of the group with LDL <70 mg/dL. According to this analysis, the hazard ratio in the “LDL 130” group is 70.6% lower (hazard ratio 0.294) than that of the “LDL 0” group (95% CI hazard ratio 0.152, 0.568).
* Slope for LDL 160: This parameter is the hazard ratio of death in the group with LDL = 160-189 mg/dL compared to that of the group with LDL <70 mg/dL. According to this analysis, the hazard ratio in the “LDL 160” group is 74.3% lower (hazard ratio 0.257) than that of the “LDL 0” group (95% CI hazard ratio 0.113, 0.580).
* Slope for LDL 190: This parameter is the hazard ratio of death in the group with LDL > 190 mg/dL compared to that of the group with LDL <70 mg/dL. According to this analysis, the hazard ratio in the “LDL 70” group is 68.3% lower (hazard ratio 0.317) than that of the “LDL 0” group (95% CI hazard ratio 0.101, 0.989).
* Intercept: This parameter corresponds to the baseline hazard function, which in this case would be the hazard of death in the group with LDL <70 mg/dL. Not naturally included in the output by Stata for a proportional hazards regression as used here.

Total: 5
	1. What analysis would you perform to assess whether the regression model used in this problem provides a “better fit” than does a model that uses only a continuous linear term for LDL? What is the result of such an analysis?

To test for linearity in the model I chose a Wald test since the proportional hazards regression was performed with robust standard error estimates. Since this analysis used dummy variables that fit a linear step function to the data, it is appropriate to use a test for linearity among all dummy variables. Using the “testparm” command in Stata, to allow for analysis of a variable list parameter for the dummy variable, results in a p-value of 0.0087, providing evidence that the data follow a nonlinear trend and a model using only a linear term would be unlikely to provide a better fit.

Did not mention including linear term (-1)

Did not mention about the test that regression coefficients for the dummy variables were 0

(-1)
Wrong p-value (-1)
Wrong conclusion (-1)

Total: 1

* 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.

Done

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: Since the variable for observation time (time to death or study termination) is censored, descriptive statistics were generated using Kaplan-Meier techniques to summarize the data according to the median survival time, minimum, maximum, the probability of surviving two years, and the probability of surviving five years described for groups according to the Mayo Clinic guidelines for LDL risk (LDL < 70 mg/dL; 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and ≥190 mg/dL). The Kaplan-Meier survival curves were displayed on a graph according to LDL category. In addition, descriptive statistics to compare the sample characteristics in each LDL group were calculated, including mean LDL measurement within the grouped range, mean age, sex, and number of deaths.

For inferential statistical analysis, the instantaneous risk (hazard) of death during the observation period was compared across groups defined by linear splines fit to LDL categories by using proportional hazards regression with robust standard error estimation (Huber-White sandwich estimator). Linear splines were fit to the LDL data with knots at the cutoff point for each LDL group (i.e. 70 mg/dL, 100 mg/dL, etc.). Parameter estimates were determined by maximum likelihood estimation from the regression analysis, with Wald-based 95% confidence intervals (CI) and two-sided p-values reported for each parameter estimate. Subjects with missing serum LDL measurements were omitted from this analysis. Statistical significance was defined by the threshold of 0.05.

Descriptive Statistics: Of a total of 735 study subjects, serum LDL measurements were available for 725. Of the 10 subjects with missing LDL measurements, two died during the study period, and the remaining 8 were still alive at study termination. Table 3 lists the descriptive statistics of interest for the subjects with complete LDL data in this study.

Table 3. Demographic information for subjects according to measured serum LDL category

|  |  |
| --- | --- |
|  | **Baseline Serum LDL (mg/dL)** |
|  | **11-69** **(n=22)** | **70-99** **(n=143)** | **100-129** **(n=228)** | **130-159** **(n=225)** | **160-189** **(n=83)** | **190-247** **(n=24)** | **All Subjects1** **(n=725)** |
| **Deaths, n** | 10 | 28 | 44 | 34 | 11 | 7 | 131 |
| **Median survival, years (95% CI)** | 5.23 (5.13, 5.57) | 5.16 (5.12, 5.57) | 5.18 (5.14, 5.23) | 5.20 (5.17, 5.26) | 5.16 (5.13, 5.23) | 5.58 (5.15. 5.74) | 5.19 (5.17, 5.20) |
| **Maximum observation time, years** | 5.75 | 5.88 | 5.88 | 5.91 | 5.91 | 5.91 | 5.91 |
| **2-year Survival, probability** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 95.4% |
| **5-year Survival, probability** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 83.5% |
| **Mean LDL, mg/dL (SD)** | 56.2 (13.6) | 86.7 (8.25) | 114.7 (8.36) | 142.7 (8.53) | 172.3 (9.21) | 208.3 (13.5) | 125.8 (33.6) |
| **Mean age, years (SD)** | 75.5 (5.63) | 74.7 (5.49) | 74.6 (5.08) | 74.2 (5.62) | 74.6 (5.67) | 76.0 (6.11) | 74.6 (5.45) |
| **Male** | 68.2% | 53.9% | 53.9% | 43.1% | 48.2% | 20.8% | 49.7% |

1Only subjects with available serum LDL measurements were included in the study

Inference: Of 735 subjects in this study, data regarding serum LDL measurements were available for 725 subjects. By proportional hazards regression modeled with linear splines fitted to serum LDL groups according to the Mayo Clinic risk categories. Among subjects with serum LDL <70 mg/dL, for each 1 mg/dL increase in serum LDL the hazard of death is 2.19% lower (hazard ratio 0.9781), which is consistent with a true population hazard within the 95% CI from 0.037% to 3.98% lower than an LDL measurement 1 mg/dL lower. For subjects with serum LDL 70-99 mg/dL, for each 1 mg/dL increase in serum LDL, the hazard of death is 2.03% lower (hazard ratio 0.9797), which is consistent with a true population hazard within the 95% CI from 4.65% lower to 0.670% higher than an LDL measurement 1 mg/dL lower. For subjects with serum LDL 100-129 mg/dL, for each 1 mg/dL increase in serum LDL, the hazard of death is 0.229% lower (hazard ratio 0.9977), which is consistent with a true population hazard within the 95% CI from 2.36% lower to 1.95% higher than an LDL measurement 1 mg/dL lower. For subjects with serum LDL 130-159 mg/dL, for each 1 mg/dL increase in serum LDL, the hazard of death is 0.361% higher (hazard ratio 1.0036), which is consistent with a true population hazard within the 95% CI from 2.06% lower to 2.84% higher than an LDL measurement 1 mg/dL lower. For subjects with serum LDL 160-189 mg/dL, for each 1 mg/dL increase in serum LDL, the hazard of death is 2.91% lower (hazard ratio 0.9709), which is consistent with a true population hazard within the 95% CI from 7.02% lower to 1.38% higher than an LDL measurement 1 mg/dL lower. For subjects with serum LDL greater than 190 mg/dL, for each 1 mg/dL increase in serum LDL, the hazard of death is 2.88% higher (hazard ratio 1.0288), which is consistent with a true population hazard within the 95% CI from 2.09% lower to 8.10% higher than an LDL measurement 1 mg/dL lower. The association between serum LDL categories at baseline and 5-year all-cause mortality is statistically significant (two-sided p<0.0001). This suggests we have sufficient evidence to reject the null hypothesis that the hazard of death is the same across LDL groups. This also suggests evidence for nonlinearity in the data.

* 1. Provide an interpretation for each parameter in your regression model, including the intercept.
* Slope for LDL0: This parameter is the incremental hazard ratio of death for each 1 mg/dL increase in serum LDL compared to the increment below among subjects with serum LDL <70 mg/dL.
* Slope for LDL70: This parameter is the incremental hazard ratio of death for each 1 mg/dL increase in serum LDL compared to the increment below among subjects with serum LDL 70-99 mg/dL.
* Slope for LDL100: This parameter is the incremental hazard ratio of death for each 1 mg/dL increase in serum LDL compared to the increment below among subjects with serum LDL 100-129 mg/dL.
* Slope for LDL130: This parameter is the incremental hazard ratio of death for each 1 mg/dL increase in serum LDL compared to the increment below among subjects with serum LDL 130-159 mg/dL.
* Slope for LDL160: This parameter is the incremental hazard ratio of death for each 1 mg/dL increase in serum LDL compared to the increment below among subjects with serum LDL 160-189 mg/dL.
* Slope for LDL190: This parameter is the incremental hazard ratio of death for each 1 mg/dL increase in serum LDL compared to the increment below among subjects with serum LDL >190 mg/dL.
* Intercept: This parameter corresponds to the baseline hazard function, which in this case would be the hazard of death in the group with LDL = 0 mg/dL. Not naturally included in the output by Stata for a proportional hazards regression as used here.
	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 test for linearity in the model I chose a Wald test since the proportional hazards regression was performed with robust standard error estimates. Since this analysis used linear splines, and a linear fit is simply a special case of linear spline models, the basic test for linearity is appropriate. Using the “test” command in Stata, analysis of all splines simultaneously results in a p-value <0.0001, providing strong evidence that the data follow a nonlinear trend and a model using only a linear term would be unlikely to provide a better fit.

* 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.

Done

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?

In Homework 1, the analyses were based primarily on point estimates and comparison between point estimates such as mean LDL or median survival time. This approach is useful for getting a quick snapshot of the big picture, and to get a sense of the sample distribution through descriptive statistics, but lends very little meaningful insight into data trends, particularly for continuous variables. Since serum LDL is a continuous measurement, and observation time (time to death or study termination) is a continuous censored variable, the point-estimation methods lose precision through dichotomization or other means of simplifying the continuous data into estimates such as means. Homework 2 and 3 used regression techniques, but the data for serum LDL was dichotomized around an “arbitrary” threshold of 160 mg/dL, requiring an assumption that all subjects with serum LDL below 160 mg/dL were similar, and likewise all subjects with serum LDL above 160 mg/dL were similar, in order to make useful inference. Information is lost by dichotomizing, and while dichotomizing often makes sense when there are clinically meaningful categories, with this dichotomized data it is difficult to identify trends and to state with confidence what information the data are really providing about the sample. Some of the advantages in using the regression strategies introduced in Homework 4 and 5 include a more natural treatment of the serum LDL variable. In these regression methods, LDL is treated as a continuous variable, and inference can be made on a population with any measured serum LDL value within the range of data obtained in the sample. For example, we can identify the increase or decrease in risk of death per unit increase of serum LDL. In this way, we have borrowed information across all subjects to infer the overall trend in 5-year all-cause mortality as it relates to serum LDL. In Homework 5 we divided serum LDL into categories, but they were clinically meaningful and based on the Mayo Clinic risk categories that are used by healthcare practitioners. Therefore, the inference can be more directly related to the information that clinicians and patients are interested in. Also, with the techniques in Homework 5, although subjects were categorized by serum LDL, the LDL variable is still treated as continuous and we are able to provide inference across all measures of serum LDL. Finally, using regression techniques we are able to detect nonlinearity more easily in the data, and can more easily detect trends (linear, curvilinear, etc.).

* 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 1 illustrates the relative hazard determined by the dummy variable method in problem 2, and the spline method in problem 3.



Figure 1. Plot of the fitted hazard ratio using a dummy variable for categories of serum LDL, or linear spline fit for the same categories of LDL. All values are relative to the hazard ratio of an individual with serum LDL = 160 mg/dL.

The method using dummy variables uses a step function to predict the trend in the data. The steps are horizontal according to the groups of serum LDL defined by the Mayo Clinic risk categories. This assumes that all individuals in a group have the same relative risk of death. Obviously, this step function loses precision in the jumps between step intervals, although the general trend of the data is captured. The method using a linear spline fit is analyzed according to the same LDL categories, but since the splines reflect the slope between “knots” the result is a more smooth fit to the data and a more clear illustration of the trend. Less information is lost since there are no jumps in relative risk, and the changes in the relative hazard for each 1 mg/dL increase in serum LDL are better captured.

* 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?

Of all the analyses to identify an unadjusted association between serum LDL and 5-year all-cause mortality, I would most prefer one of the regression methods used in Homework 5. Specifically, I would choose to conduct an analysis with proportional hazards regression using a linear spline fit of the continuous LDL variable according to the clinically meaningful groups defined by the Mayo Clinic risk categories. By using regression, there is minimal loss of data due to dichotomization, although by categorizing the LDL data there is likely some loss of precision. This method also allows for “borrowing” information across groups to give reasonable inference of risk of death for all serum LDL values in the data range. Finally, this method allows for more nuanced identification of trends across data, including nonlinear trends, and is more robust at estimating linearity.