**5498**

**Biost 515 (Winter 2014)**

**Instructor: Scott Emerson**

**Homework 5**

*Problems 2 and 3 of the homework build on the analyses performed in homeworks #1 through #4. As such, all questions relate to associations among death from any cause, serum low density lipoprotein (LDL) levels, age, and sex in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. See homework #1 for additional information. Problem 1 of this homework uses the same dataset to explore associations between prevalence of diabetes and race in the population from which that sample was drawn.*

1. Perform a statistical regression analysis evaluating an association between prevalence of diabetes and race by comparing the odds of a diabetes diagnosis across.
   1. Fit a logistic regression model that uses whites as a reference group. Is this a saturated model? Provide a formal report (methods and inference) about the scientific question regarding an association between diabetes and race.

**Methods:** The odds of a diabetes diagnosis were compared between subjects with varying race using a logistic regression model. Three categories of race and a fourth “all other” category were defined in this analysis: whites, blacks, Asians, and other. The white group was used as the reference group, and therefore represents the model intercept; the remaining three groups were modeled as indicator variables. Therefore, the logistic regression model contains four groups with three predictors plus an intercept, so the model is saturated and parameter estimates agree exactly with sample measurements. Quantification of association between diabetes and race was tested by simultaneously testing each term had coefficients equal to one and was based on the Chi-Squared test statistic with 3 degrees of freedom. A two-sided p-value was computed using the approximate normal distribution for logistic regression parameter estimates, and significance was determined at the 0.05 level.

**Inference:** Data was available on 735 subjects, with race defined into four groups: white (n=572), black (n=104), Asian (n=47), and other (n=12). It deserves mention that we cannot draw strong scientific meaning from the group of subjects categorized as other since we cannot assume the group is homogenous in another race. As illustrated in Table 1, black subjects had the highest proportion of diabetes diagnoses (17.31%), while 9.79% of white subjects, 6.38% of Asians, and 16.67% of other subjects were diagnosed with diabetes. The sample odds of a diabetes diagnosis was 0.209 among black subjects, 0.109 among white subjects, 0.068 among Asian subjects, and 0.200 among all other subjects.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 1: Baseline Diabetes Characteristics of the Study Population | | | | | |
|  | Race | | | |  |
| Diabetes | White  (n = 572) | Black  (n = 104) | Asian  (n = 47) | Other  (n = 12) | Overall  (n = 735) |
| N  Prevalence  Odds | 56  0.098  0.109 | 18  0.173  0.209 | 3  0.064  0.068 | 2  0.167  0.200 | 79  0.107  0.120 |
|  | | | | | |

Based on a Chi-squared test with three degrees of freedom, a two-sided p-value of 0.1096 suggests that we do not have the statistical significance necessary to reject the null hypothesis that the odds of a diabetes diagnosis are not associated with race. (Note: since the Chi-squared test led to insignificant results, no pairwise comparison between race parameters need mention.)

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

**Answer:** Using the regression model fit in part (a), the intercept β0 = -0.221 can interpreted as the odds of white adults aged 65 years and older having diabetes are e-0.221 = 0.109. The parameter β1 = 0.657 can be interpreted as the odds of diabetes in elderly black adults are e0.657 = 1.929 times greater than the odds of diabetes in elderly white adults. The parameter β2 = -0.465 can be interpreted as the odds of diabetes in elderly Asian adults are e-0.465 = 0.628 times less than the odds of diabetes in elderly white adults. The parameter β3 = 0.611 can be interpreted as the odds of diabetes in elderly adults who are neither black nor Asian are e0.611 = 1.843 times greater than the odds of diabetes in elderly white adults.

* 1. If we were to ignore issues 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:** If we were to ignore issues related to multiple comparisons, we could conclude that the observed odds ratio of 1.929 for the black group in comparison to the white group would not be judged unusual if the true odds ratio were anywhere between 1.082 and 3.438 (p-value = 0.026). Similarly, the observed odds ratio of 0.628 for the Asian group compared to the white group would not seem unusual if the true odds ratio were anywhere between 0.189 and 2.089 (p-value = 0.448); and the observed odds ratio of 1.843 for the all-other group compared to the white group would not seem unusual if the true ratio were between 0.394 and 8.622 (p-value = 0.437). Therefore, using a 0.05 level of significance, we would reject the null hypothesis of no association between the diagnosis of diabetes and race for an alternative hypothesis that the odds of diabetes are higher in elderly black adults than in elderly white adults.

* 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)?

**Answer:** If we fit a logistic model that uses blacks as a reference group, our new model will be a reparameterized version of the model in part (a). As such, we draw the same inference from part (a): a two-sided p-value of 0.1096 suggests that we do not have the statistical significance to reject the null hypothesis that the odds of a diabetes diagnosis are not associated with race.

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

**Answer:** Using the logistic regression model fit in part (d), the intercept β0 = -1.564 can interpreted as the odds of black adults aged 65 years and older having diabetes are e-1.564 = 0.209. The parameter β1 = -0.657 can be interpreted as the odds of diabetes in elderly white adults are e-0.657 = 0.519 times less than the odds of diabetes in elderly black adults. The parameter β2 = -1.122 can be interpreted as the odds of diabetes in elderly Asian adults are e-1.122 = 0.326 times less than the odds of diabetes in elderly black adults. The parameter β3 = -0.045 can be interpreted as the odds of diabetes in elderly adults who are neither white nor Asian are e-0.045 = 0.9556 times less than the odds of diabetes in elderly black adults.

* 1. If we were to ignore issues 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:** If we were to ignore issues related to multiple comparisons, we could conclude that the observed odds ratio of 0.519 for the white group in comparison to the black group would not be judged unusual if the true odds ratio were anywhere between 0.291 and 0.924 (p-value = 0.026). Similarly, the observed odds ratio of 0.326 for the Asian group compared to the black group would not seem unusual if the true odds ratio were anywhere between 0.091 and 1.166 (p-value = 0.085); and the observed odds ratio of 0.956 for the all-other group compared to the black group would not seem unusual if the true ratio were between 0.193 and 4.737 (p-value = 0.956). Therefore, using a 0.05 level of significance, we would reject the null hypothesis of no association between the diagnosis of diabetes and race for an alternative hypothesis that the odds of diabetes are higher in elderly black adults than in elderly white adults.

* 1. What do your results from parts (c) and (f) say about the dangers of using the p values for individual regression parameters from a dummy variable regression to decide whether to include or exclude those variables in a regression model (i.e., in a “stepwise model building” procedure)?

**Answer:** In parts (c) and (f), statistical significance was judged at the 0.05 level and there were no differences in the decision on whether or not to reject each of the three odds ratio pairwise comparisons. However, had our statistical significance been determined at the 0.10 level, we would have concluded from model (c) that there is no statistically significant difference between the odds of diabetes in Asians compared to whites, and we would have concluded from model (f) that there is a statistically significant difference between the odds of diabetes in Asians compared to blacks. Had we removed Asians from the model in part (c), we would never have gained the inference from part (f). Therefore, p-values for individual regression parameters from a dummy variable regression should never be used to determine inclusion/exclusion in a 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

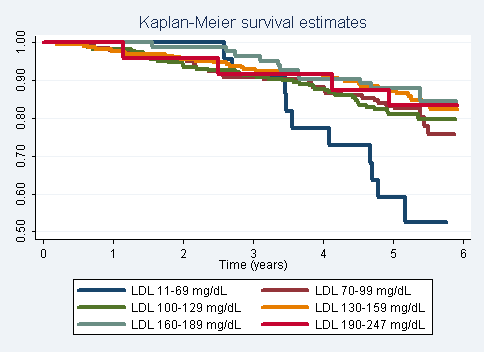
egen ldlCTG = cut(ldl), at(0 70 100 130 160 190 250)

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

**Methods:** In this study, descriptive statistics are presented for serum LDL levels as well as the censoring distribution of all-cause mortality. For the purposes of the analysis, serum LDL (measured in mg/dL) was categorized according to Mayo Clinic guidelines: LDL ≤ 69 (ideal for people at very high risk of heart disease), 70 ≤ LDL ≤ 99 (ideal for people at risk of heart disease), 100 ≤ LDL ≤ 129 (near ideal), 130 ≤ LDL ≤ 159 (borderline high), 160 ≤ LDL ≤ 189 (high), and LDL ≥ 190 (very high). Within each category, Kaplan-Meier survival estimates were calculated and graphed.

Proportional hazards regression modeling serum LDL as dummy variables was used to assess the association between all-cause time to death and groups defined by baseline serum LDL. Quantification of association between all-cause mortality and serum LDL was summarized by hazard ratios and tested by simultaneously testing that all terms had coefficients equal to one. Two sided p-values for this Chi-squared five degree of freedom test were computed using Wald Statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of the study accrual were omitted from the analysis.

**Inference:** This study consisted of 735 subjects, for which a Kaplan-Meier estimated average follow-up time of 5.33 years (range: 5.00 – 5.91 years) was calculated. Data was missing for 10 of the subjects, and as such, these subjects are omitted from all analyses. We cannot assess the impact that such omissions might have on the generalizability of our results. In the remaining 725 subjects, the mean LDL was 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL) and 131 subjects were observed to die during the study. The following graph illustrates Kaplan-Meier survival estimates for each of the six serum LDL strata recommended by the Mayo Clinic. Table 2 presents Kaplan-Meier estimates of distribution of time from study enrollment to all-cause death, according to the six serum LDL strata as well as for the entire sample. At five years of follow-up, the probability of survival among participants still alive was highest for those with LDL between 160 mg/dL - 189 mg/dL and lowest for those with LDL ≤ 59.1 mg/dL.



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Kaplan-Meier Survival Estimates | | | | | | | |
|  | Serum LDL (mg/dL) at Study Enrollment | | | | | |  |
|  | 11-69 | 70-99 | 100-129 | 130-159 | 160-189 | 190-247 | Overall1 |
| N Subjects | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| N Deaths | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| Survival Probability (%) |  |  |  |  |  |  |  |
| 1 year | 100 | 97.9 | 98.3 | 97.8 | 100 | 100 | 98.1 |
| 2 year | 100 | 95.8 | 93.9 | 95.6 | 98.8 | 95.8 | 95.4 |
| 3 year | 90.9 | 90.9 | 91.2 | 92.9 | 96.4 | 91.7 | 92.1 |
| 4 year | 77.3 | 88.1 | 87.7 | 91.1 | 90.4 | 91.7 | 88.8 |
| 5 year | 59.1 | 83.2 | 81.1 | 87.1 | 88.0 | 83.3 | 83.5 |
| 1Among subjects with serum LDL data available | | | | | | | |

Based on a Chi-squared test with five degrees of freedom, a two-sided p-value of 0.0087 suggests that we have sufficient statistical evidence to reject the null hypothesis of no association between serum LDL levels and all-cause mortality in favor of a hypothesis that there is an association between all-cause mortality and serum LDL dummy variables modeled according to Mayo Clinic guidelines.  
  
3/3 for descriptive statistics

3/3 for performing an appropriate analysis

1/4 for reporting the association appropriately

Wrong interpretation of coefficient (-2)

Wrong interpretation of CI (-1)

Total: 7

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

**Answer:** Using the model fit in part (a), the intercept β0 can be interpreted as the baseline hazard function, which need not be estimated in order to estimate regression parameters. The parameter β1 = 0.398 implies that the instantaneous risk of death is a relative 60.2% lower among groups with LDL anywhere between 70-99 mg/dL than groups with LDL < 70 mg/dL. The parameter β2 = 0.393 implies that the instantaneous risk of death is a relative 60.7% lower among groups with LDL anywhere between 100-129 mg/dL than groups with LDL < 70 mg/dL. The parameter β3 = 0.294 implies that the instantaneous risk of death is a relative 70.6% lower among groups with LDL anywhere between 130-159 mg/dL than groups with LDL < 70 mg/dL. The parameter β4 = 0.257 implies that the instantaneous risk of death is a relative 74.3% lower among groups with LDL anywhere between 160-189 mg/dL than groups with LDL < 70 mg/dL. And finally, the parameter β5 = 0.317 implies that the instantaneous risk of death is a relative 68.3% lower among groups with LDL ≥ 190 mg/dL than groups with LDL < 70 mg/dL.   
  
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?

**Answer:** 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 serum LDL, we can perform proportional hazards regression modeling LDL as dummy variables and adding a linear LDL term. We can then perform a test on the dummy variables and determine their significance (i.e. test for a straight line). Based on this sort of analysis, a Chi-squared test with five degrees of freedom suggests that we do not have sufficient evidence to reject the null hypothesis that LDL is a nonlinear term (p-value = 0.3988).  
  
Total: 5

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.

**Answer:** *(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.

**See problem 2 for methods and report of descriptive statistics.**

**Methods:** Proportional hazards regression modeling serum LDL as a piecewise linear variable was used to assess the association between all-cause time to death and groups defined by baseline serum LDL. The intervals for which LDL was modeled linearly are those defined in problem 2 as suggested by the Mayo Clinic. Quantification of association between all-cause mortality and serum LDL was summarized by hazard ratios and tested by simultaneously testing that all terms had equal coefficients. Two sided p-values for this Chi-squared six degree of freedom test were computed using Wald Statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of the study accrual were omitted from the analysis.

**Inference:** Based on a Chi-squared test with six degrees of freedom, a two-sided p-value <0.001 strongly suggests that we have sufficient statistical evidence to reject the null hypothesis of no association between serum LDL levels and all-cause mortality in favor of a hypothesis that higher levels of serum LDL are associated with longer survival.

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

**Answer:** Using the model fit in part (a), the intercept β0 can be interpreted as the instantaneous risk of death when serum LDL equals zero. The parameter β1 = 0.978 implies that for every 1 mg/dL increase in serum LDL, individuals with serum LDL ≤ 69 mg/dL are at 0.978 times lower risk of instantaneous death. The parameter β2 = 0.980 implies that for every 1 mg/dL increase in serum LDL, individuals with serum LDL between 70 and 99 mg/dL are at 0.980 times lower risk of instantaneous death. The parameter β3 = 0.998 implies that for every 1 mg/dL increase in serum LDL, individuals with serum LDL between 100 and 129 mg/dL are at 0.998 times lower risk of instantaneous death. The parameter β4 = 1.004 implies that for every 1 mg/dL increase in serum LDL, individuals with serum LDL between 130 and 159 mg/dL are at 1.004 times higher risk of instantaneous death. The parameter β5 = 0.971 implies that for every 1 mg/dL increase in serum LDL, individuals with serum LDL between 160 and 189 mg/dL are at 0.971 times lower risk of instantaneous death. The parameter β6 = 1.029 implies that for every 1 mg/dL increase in serum LDL, individuals with serum LDL ≥ 190 mg/dL are at 1.029 times higher risk of instantaneous death.

* 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?

**Answer:** 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 serum LDL, we can perform proportional hazards regression modeling LDL with linear splines and adding a linear LDL term. We can then perform a test on the splines variables and determine their significance (i.e. test for a straight line). Based on this sort of analysis, a Chi-squared test with six degrees of freedom suggests that we do not have sufficient evidence to reject the null hypothesis that LDL is a nonlinear term (p-value = 0.0788).

* 1. For each population defined by serum LDL value, compute the hazard ratio relative to a group having serum LDL of 160 mg/dL. (This will be used in problem 4). This can be effected by generating fitted hazard ratio estimates for each individual in the sample, and then dividing that fitted value by the fitted value for a subject having a LDL of 160 mg/dL.

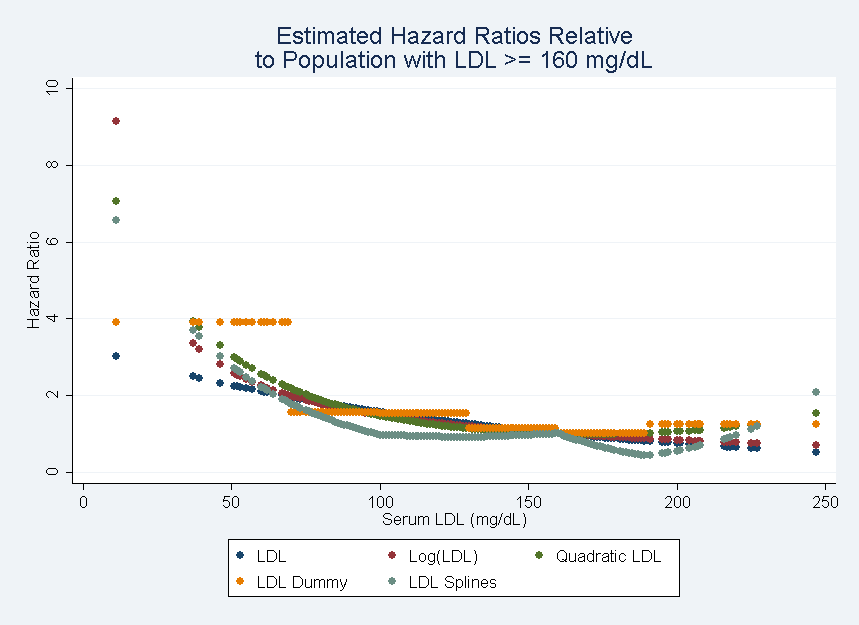
**Answer:** *(done)*

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?

**Answer:** The regression strategies used in Homeworks 4 and 5 treat serum LDL as a continuous variable and are more advantageous than those in 1 through 3, for which serum LDL was dichotomized and therefore suffered from loss of precision. Furthermore, the analyses in Homework 4 and 5 condition on the variable which we think of as the predictor (i.e. serum LDL), whereas Homework 1 through 3 condition on the variable which we consider the outcome (i.e. mortality). Conditioning on the predictor of interest seems more intuitive and scientifically 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.

**Answer:** The following graph displays the fitted values from each of the 5 models fit in Homeworks 4 and 5. All five models predict the instantaneous risk of death is greatest for extreme values of serum LDL (with lower LDL being more hazardous than higher LDL). However, since less information is borrowed across intervals when treating LDL as dummy variables or linear splines variables, smooth (U-shaped) curves are not as easily interpolated in areas for which less data is available.



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

**Answer:** A priori, I would prefer the analysis performed in homework 4 in which we treated LDL as a continuous linear variable to explore an association between all-cause mortality and serum LDL in an elderly population. The benefits to this sort of analysis are that (1) we do not lose any precision by dichotomizing a continuous variable, (2) we are able to adjust for any potential confounding much easier than if we treated LDL as dummy or linear splines variables, (3) it is much easier to draw scientific interpretation and meaning from model parameters and results, and (4) we are conditioning on the predictor of interest rather than the outcome of interest, which is more scientifically pleasing.