**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #6**

March 4, 2015

Problems 1-3 of the homework relate to the dataset regarding MRI measurements of cerebral atrophy in elderly Americans (mri.doc and mri.txt). In this homework we will focus primarily on associations between mortality and serum LDL as possibly modified by race.

1. Suppose we are interested in exploring whether any association between time to death and serum LDL is adequately modeled by a relationship in which the log hazard function is linear in LDL. I ask you to compare several different alternative models that allow nonlinearity. In part f, I ask you to plot fitted HR estimates from each of these models on the same axis. In order to have comparability across models, we need to use the same reference group:

* In all parts of this problem where you need to divide the LDL values into intervals, use 70, 100, 130, and 160 mg/dL as breakpoints for the LDL measurements. Stata commands that might be used are:

egen ldlctg= cut(ldl), at(0,70,100,130,160,400)

mkspline sldlA 70 sldlB 100 sldlC 130 sldlD 160 sldlE = ldl

* In all parts of this problem where you model LDL continuously, we will use 1 mg/dL as the reference group (this will accommodate the log transformation). Thus you might create variables in Stata:

g logldl= log(ldl)

g cldl= ldl – 1

g cldlsqr= cldl^2

g cldlcub= cldl^3

* 1. Fit a regression model in which you test for a linear relationship using a step function as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum LDL level and mortality rate. LDL values were divided into intervals using 70, 100, 130, and 160 mg/dL as breakpoints. Distributions of time to death from any cause were modeled on the serum LDL level as an untransformed term and dummy variables (the group with less than 70mg/dL LDL as the reference group). Serum LDL level as dummy variables were used to test for linearity and the two-side P value was determined based on Wald statistics.

**Results:** There are a total of 735 observations in the study and 10 observations have missing values on their LDL reading. These ten observations have been removed from the data set. When testing whether the coefficients are zero for the serum LDL as dummy variables, the two-sided p value from the overall F test is 0.361. Therefore, we cannot reject the null hypothesis that there were no departures from linearity in the model that time to death from any cause were modeled on the serum LDL level.

* 1. Fit a regression model in which you test for a linear relationship using a quadratic polynomial as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum LDL level and mortality rate. Distributions of time to death from any cause were modeled on the serum LDL level in two terms: an untransformed term and a square term. The two-side P value was determined based on Wald statistics.

**Results:** There are a total of 735 observations in the study and 10 observations have missing values on their LDL reading. These ten observations have been removed from the data set. From proportional hazards regression analysis, the two-sided p value for the slope of the square term is 0.0552. Therefore, we cannot reject the null hypothesis that there were no departures from linearity in the model that time to death from any cause were modeled on the serum LDL level using a quadratic polynomial as an alternative model.

* 1. Fit a regression model in which you test for a linear relationship using a cubic polynomial as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum LDL level and mortality rate. Distributions of time to death from any cause were modeled on the serum LDL level in three terms: an untransformed term, a square term and a cube term. The square term and cube term related to serum LDL level were used to test for linearity and the two-side P value was determined based on Wald statistics.

**Results:** There are a total of 735 observations in the study and 10 observations have missing values on their LDL reading. These ten observations have been removed from the data set. When testing whether the coefficients are zero for the square term and the cube term, the two-sided p value from the overall F test is 0.0164. Therefore, we reject the null hypothesis that there were no departures from linearity in the model that time to death from any cause were modeled on the serum LDL level using a quadratic polynomial as an alternative model.

* 1. Fit a regression model in which you test for a linear relationship using linear splines as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum LDL level and mortality rate. Distributions of time to death from any cause were modeled on the serum LDL level using linear splines as an alternative model. LDL values were divided into intervals using 70, 100, 130, and 160 mg/dL as breakpoints. These spline terms related to serum LDL level were used to test for linearity and the two-side P value was determined based on Wald statistics.

**Results:** There are a total of 735 observations in the study and 10 observations have missing values on their LDL reading. These ten observations have been removed from the data set. When testing whether the coefficients are equal for all intervals, the two-sided p value from the overall F test is 0.119. Therefore, we cannot reject the null hypothesis that there were no departures from linearity in the model that time to death from any cause were modeled on the serum LDL level using linear splines as an alternative model.

* 1. Fit a regression model in which you test for a linear relationship using a logarithmic transformation as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

**Method:** A propotional hazard regression analysis using Huber-White estimates of the standard error was performed to evaluate the association between serum LDL level and mortality rate. Distributions of time to death from any cause were modeled on the serum LDL level in two terms: an untransformed term and a logarithmic transformed term. The two-side P value was determined based on Wald statistics.

**Results:** There are a total of 735 observations in the study and 10 observations have missing values on their LDL reading. These ten observations have been removed from the data set. From proportional hazards regression analysis, the two-sided p value for the slope of the logarithmic transformed term is 0.004. Therefore, we reject the null hypothesis that there were no departures from linearity in the model that time to death from any cause were modeled on the serum LDL level using a logarithmic transformation as an alternative model.

* 1. On the same set of axes, plot the fitted values from each of the above models, as well as a model that includes only the (centered) serum LDL values. Comment on the similarity and/or differences among these models. How might these results guide your choice of a particular model when investigating whether associations are not well described by a linear relationship?

**Answer:**



The fitted values are very similar in the models using a square polynomial, cubic polynomial or the linear splines as an alternative model. However, the fitted values are different in the models using dummy variables, a logarithmic transformation or centered LDL as an alternative model.

Therefore, I would choose either polynomial function or the linear spines when investigating whether associations are not well described by a linear relationship.

1. Consider again a model exploring the associations between time to death and serum LDL when using linear splines.
   1. Explain the interpretation of the regression parameters in such a model.

**Answer:** For patients with serum LDL level less than 70mg/dL, their hazard rate drop 2.18% when their LDL level increase by 1mg/dL.

For patients with serum LDL level within the range of [70, 100) mg/dL, their hazard rate drop 2.07% when their LDL level increase by 1mg/dL.

For patients with serum LDL level within the range of [100, 130) mg/dL, their hazard rate drop 0.0908% when their LDL level increase by 1mg/dL.

For patients with serum LDL level within the range of [130, 160) mg/dL, their hazard rate drop 0.195% when their LDL level increase by 1mg/dL.

For patients with serum LDL level within the range of [160, 400) mg/dL, their hazard rate drop 0.611% when their LDL level increase by 1mg/dL.

* 1. Is there evidence that the association between time to death and serum LDL is truly U-shaped? Explain your evidence.

**Answer:** No. The coefficients for each interval are all less than 1, which indicates that there is no change in the trend. Therefore, we do not have evidence that the association between time to death and serum LDL is truly U-shaped.

1. Suppose we are interested in exploring the associations between time to death and serum LDL as possibly modified by race. In this problem you do not need to provide formal description of the methods or inference, though I do ask at times for specific inferential quantities.
   1. Fit a model of time to death regressed on a log transformation of serum LDL, race, and their interaction. Provide an explicit interpretation of each parameter in your model (be sure to include the actual numeric value in your interpretation, but you do not have to provide CI or p values for this part).

**Answer:**

Hazard ratio = β0 + β1 \* log2(ldl) + β2 \* Black + β3 \* Asian + β4 \* Other + β5 \* log2(ldl) \* Black + β6 \* log2(ldl) \* Asian + β7 \* log2(ldl) \* Other

β1: For Whites, the hazard ratio is 0.585 when comparing a high LDL group with low LDL group whose serum LDL level is half.

β2: For Blacks, the hazard ratio is 0.154 if the serum LDL level is 1mg/dL.

β3: For Asians, the hazard ratio is 305 if the serum LDL level is 1mg/dL.

β4: For other race, the hazard ratio is 3.33e+08 if the serum LDL level is 1mg/dL.

β5: When comparing Black and White, the difference in the difference of the hazard ratio is 1.36 when comparing a high LDL group with low LDL group whose serum LDL level is half.

β6: When comparing Asian and White, the difference in the difference of the hazard ratio is 0.444 when comparing a high LDL group with low LDL group whose serum LDL level is half.

β7: When comparing other race and White, the difference in the difference of the hazard ratio is 0.0616 when comparing a high LDL group with low LDL group whose serum LDL level is half.

* 1. Use the regression analysis in part a to perform a statistical test of the hypothesis that race does not modify the association between time to death and serum LDL. Make clear which parameters you test and provide a two-sided p value.

**Answer:** The coefficients for the products of the log transformed serum LDL and race as dummy variables (β5-β7) were tested whether they are all zero. The two-sided p value is 0.0452 based on Wald statistics. Therefore, we can reject the null hypothesis that race does not modify the association between time to death and serum LDL.

* 1. Use the regression analysis in part a to perform a statistical test of the hypothesis that there is no association between time to death and serum LDL. Make clear which parameters you test and provide a two-sided p value.

**Answer:** The coefficients for all terms containing the log transformed serum LDL (β1, β5-β7) were tested whether they are all zero. The two-sided p value is less than 0.00001 based on Wald statistics. Therefore, we can reject the null hypothesis that there is no association between time to death and serum LDL.

* 1. Use the regression analysis in part a to perform a statistical test of the hypothesis that there is no association between time to death and race. Make clear which parameters you test and provide a two-sided p value.

**Answer:** The coefficients for all terms containing race as dummy variables (β2-β7) were tested whether they are all zero. The two-sided p value is less than 0.00001 based on Wald statistics. Therefore, we can reject the null hypothesis that there is no association between time to death and race.

* 1. Use the regression analysis in part a to perform a statistical test of the hypothesis that there is no difference in the distribution of time to death between whites and blacks. Make clear which parameters you test and provide a two-sided p value.

**Answer:** The coefficients for all terms containing race of blacks in the dummy variables with the race group of whites as reference group (β2 and β5) were tested whether they are all zero. The two-sided p value is 0.542 based on Wald statistics. Therefore, we can not reject the null hypothesis that there is no difference in the distribution of time to death between whites and blacks.

Problems 4 of the homework relates to the university salary dataset.

1. We are interested in raises given to faculty hired in recent years. For this problem, restrict attention to faculty hired in 1990 or later and who started at the university within one year of the year in which they received their highest degree. In order to (at least in part) examine the patterns of raises given to faculty, we will model salaries by sex, calendar year, and an interaction between sex and calendar year. Use such a model to answer the following questions.
   1. Is there evidence of sex discrimination in the mean salary given in recent years? You do not have to provide full inference, but you should make clear the basis for your answer.

**Answer:** In the model with the mean salary after clustering, coefficients for two parameters (sex, the product of sex and calendar year) were tested whether they are all zero. The two-sided p value from overall F test is 0.0493. Therefore, we can reject the null hypothesis that there is no association between sex and the mean salary given in recent years.

* 1. Is there evidence of sex discrimination in the geometric mean salary given in recent years? You do not have to provide full inference, but you should make clear the basis for your answer.

**Answer:** In the model with the geometric mean salary after clustering, coefficients for two parameters (sex, the product of sex and calendar year) were tested whether they are all zero. The two-sided p value from overall F test is 0.0216. Therefore, we can reject the null hypothesis that there is no association between sex and the mean salary given in recent years

* 1. What are the relative merits of the two models used in parts a and b?

**Answer:** The model based on the mean salary is easy to interpret and compare. The model based on the geometric mean can be useful when the response is multiplicative and it can also downweight the outliers in the data and gives

* 1. If you answered parts a and b correctly, you accounted for the correlated observations used in the analysis. Compare that inference to what you would have obtained had you incorrectly treated the data as independent. In particular, consider whether these incorrect models would have tended to be conservative or anti-conservative when making inference about associations with sex. How would your answer differ when considering associations by year?

**Answer:** If clustering is not considered in the model, coefficients for two parameters (sex, the product of sex and calendar year) were tested whether they are all zero. The p-values from the overall F test are less than 0.00001 in both models. The p-values are becoming smaller, so the results tend to be anti-conservative when the data is treated as independent.

When answering the associations by year, coefficients for two parameters (sex, the product of sex and calendar year) were tested whether they are all zero. The p-values from the overall F test are less than 0.00001 no matter whether clustering is considered. However, the F test results lower when clustering is included in the model. Therefore, the results tend to be conservative when the data is treated as independent.