**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #6**

March 4, 2015

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Wednesday, March 11, 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***

1. ***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.***
2. ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

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).

**Answer:**

**I used step function to test linearity by including linear term and test for the categorical variable.**

**The model takes the form:**

**“cldl” is the continuous term of ldl, and the rest of the variables are categorized ldl from level [70-100] to level [160-400].**

**The parameters used to test the null hypothesis that there were no departures from linearity are. They were tested simultaneously by a waldtest:**

**(I used waldtest in STATA for all my tests, and for Proportional hazard regressions, I used robust standard error.)**

**p-value： 0.3609 (>0.05), hence we do not have high confidence to reject the null hypothesis that .**

**To conclude, we do not have high confidence in nonlinear association between time to death and serum LDL. But notice this does not prove linearity. Nonlinearity could be in ways other than step function.**

* 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).

**Answer:**

**I used quadratic polynomial to test linearity by including linear term and test for the categorical variable.**

**The model takes the form:**

**“cldl” is the continuous term of ldl, and “cldlsqr” is the quadratic form of ldl.**

**The parameter used to test the null hypothesis that there were no departures from linearity is . Because it is the only parameter getting tested, we can simply look at the regression result to look for a two-sided p-value:**

**p-value= 0.05482 (>0.05), hence we do not have high confidence to reject the null hypothesis that .**

**To conclude, we do not have high confidence in nonlinear association between time to death and serum LDL. But notice this does not prove linearity. Nonlinearity could be in ways other than quadratic polynomial.**

* 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).

**Answer:**

**I used cubic polynomial to test linearity by including linear term and test for the categorical variable.**

**The model takes the form:**

**“cldl” is the continuous term of ldl, “cldlsqr” is the quadratic form of ldl, and “cldlcub” is the cubic form of ldl.**

**The parameter used to test the null hypothesis that there were no departures from linearity is . They were tested simultaneously by a waldtest:**

**p-value= 0.01634 (<0.05), hence we have high confidence to reject the null hypothesis that .**

**To conclude, we have high confidence in nonlinear association between time to death and serum LDL. This nonlinearity could be captured by a cubic polynomial.**

* 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).

**Answer:**

**I used linear splines to test linearity by including linear term and test for the categorical variable.**

**The model takes the form:**

**+**

**“cldl” is the continuous term of ldl, “spline” is the splines of ldl, and the knots of the splines are (70,100,130,160).**

**The parameter used to test the null hypothesis that there were no departures from linearity is . They were tested simultaneously by a waldtest:**

**p-value= 0.1191 (>0.05), hence we do not have high confidence to reject the null hypothesis that.**

**To conclude, we do not have high confidence in nonlinear association between time to death and serum LDL. But notice this does not prove linearity. Nonlinearity could be in ways other than linear splines.**

* 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).

**Answer:**

**I used log-transformed ldl to test linearity by including linear term and test for the categorical variable.**

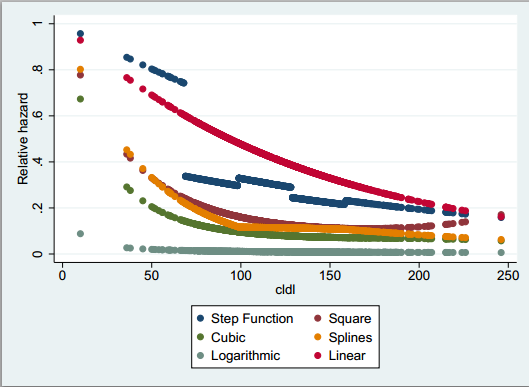
**The model takes the form:**

**“cldl” is the continuous term of ldl, “logldl” is the log-transformed ldl.The parameter used to test the null hypothesis that there were no departures from linearity is . Because it is the only parameter getting tested, we can simply look at the regression result to look for a two-sided p-value:**

**p-value= 0.0036 (<0.05), hence we have high confidence to reject the null hypothesis that.**

**To conclude, we have high confidence in nonlinear association between time to death and serum LDL. This nonlinearity could be captured by log-transformed ldl.**

* 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?



**Compared to the step function model (from which we can see that the last category with lowest ldl has the highest instantaneous risk of death), we can assume that the linear model is going to be affected (skewed) to the low ldl group. Hence we might need a more flexible modeled ldl.**

**The fitted lines given by quadratic polynomial, linear splines and cubic polynomial are similar. The linear splines also pick up the drastic downward trend in the low ldl levels.**

**The fitted line given by log-transformed ldl seem to be too flat. It may suggest that by log-transforming the ldl, the association between instantaneous risk of death and ldl is not apparent.**

**I would choose linear splines. Because it contains information given by quadratic polynomial, cubic polynomial and step function. But it takes into consideration of the ordering of the groups and does not assume smoothness. Hence it is a good choice to help us model ldl more accurately without going to far as to require more strict assumptions.**

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.

**The slopes are estimated hazard ratio between two groups differing by 1 mg/dl ldl within the sample interval defined by knots.**

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

**The estimates of are all below 1, so the trend of the association does not change, which means no sign of U-shaped association.**

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).

**.1 fold comparison between ldl groups:**

**hazard of the group with ldl=1mg/dl (which we do not concern ourselves with).**

**(w=2,3,4)**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **W=2 (black)** | **0.969** | **1.043** |
| **W=3 (asian)** | **0.831** | **0.894** |
| **W=4 (other)** | **0.633** | **0.682** |

**,**

**.**

* 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.

**To test wether there was effect modification, we have to test the interactions involving race together.**

**The two-sided p-value (given by R-waldtest) is 0.045.(<0.05), hence we have high confidence to reject the null hypothesis that there was no effect modification given by race.**

* 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.

**To test wether there was association between ldl and time to death, we have to test all the parameters related to ldl.**

**The two-sided p-value (given by R-waldtest) is <0.001, hence we have high confidence to reject the null hypothesis that there was no association between ldl and time to death.**

* 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.

**To test wether there was association between race and time to death, we have to test all the parameters related to ldl.**

**The two-sided p-value (given by R-waldtest) is <0.001, hence we have high confidence to reject the null hypothesis that there was no association between race and time to death.**

* 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.

**To test wether the distribution of time to death between whites and blacks are the same, we have to test all the parameters related to race black.**

**The two-sided p-value (given by STATA) is 0.5416, hence we do not have high confidence to reject the null hypothesis that the distribution of time to death between whites and blacks are the same.**

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.

**I used the model mean salary as response variable, sex as predictor of interest, and the interaction term was included. (We have multiple entries for each id for different years, hence we have to account for the correlation between these multiple entries for the same person. I used cluster() to accoun for correlation.)**

**By looking at the two-sided p-value given by the partial test of the coefficients related to sex(0.235>0.05), we can not reject the null hypothesis that the mean salary of male and female are the same.**

* 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.

**I used the model geometric mean salary as response variable, sex as predictor of interest, and the interaction term was included. (We have multiple entries for each id for different years, hence we have to account for the correlation between these multiple entries for the same person. I used cluster() to accoun for correlation.)**

**By looking at the two-sided p-value given by the partial test of the coefficients related to sex (0.122=3>0.05), we can not reject the null hypothesis that the geometric mean salary of male and female are the same.**

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

**Both of them accounted for correlated observations. Use the log-transformed salary (considering geometric mean) when we have prior information suggesting salary changes in a multiplicative scale (in this case we have); use the mean when we do not have strong prior indicating a multiplicative scale change in salary, and when we try to quantify difference in a more understandable summary measure.**

* 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?

**If we treat correlated observations as imdepent observations, we would fit a model without clustering for id. Here are the comparisons of the two-sided p-values:**

|  |  |  |
| --- | --- | --- |
|  | **Clustered** | **Unclustered** |
| **Mean** | **0.235** | **0.002** |
| **Geometric mean** | **0.123** | **0.0001** |

**As presented in the table, we can see that unclustered p-values are significantly smaller than clustered p-values, which indicated that when we fail to account for correlated data, we have anti-conservative results.**