**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).

I fit a cox proportional hazard model simultaneously to continuous ldl and (ldl – 1) and ldl as dummy variables at 70, 100, 130 and 160 mg/dL. Because I was using dummy variables with categorized continuous variables in a non-saturated model, in order to test linearity, I added a linear term (ldl – 1) and then tested the dummy variables together. With a two-sided p-value of 0.3609, we cannot reject the null hypothesis that all coefficients were zero (i.e., there was no departures from linearity), thus we cannot conclude that the relationship between serum LDL and log hazard function is non-linear when using dummy variables to detect nonlinearity.

* 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).

I fit a cox proportional hazard model simultaneously to continuous ldl (ldl – 1) and a quadratic polynomial as an alternative model (the square of LDL). With a two-sided p-value of 0.055, we cannot reject the null hypothesis that there were no departures from linearity, and thus cannot conclude that the relationship between serum LDL and risk of death is nonlinear.

* 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).

I fit a cox proportional hazard model simultaneously to continuous ldl (ldl – 1) and a quadratic model and cubic polynomial as an alternative model (the cube of LDL). I did a test to see if the coefficients of the quadratic and cubic polynomial were simultaneously zero. With a two-sided p-value of less than 0.0001, we can reject the null hypothesis that there were no departures from linearity in the relationship between LDL and risk of death, and thus can conclude that the relationship is non-linear and is better modeled while including the quadratic and cubic terms.

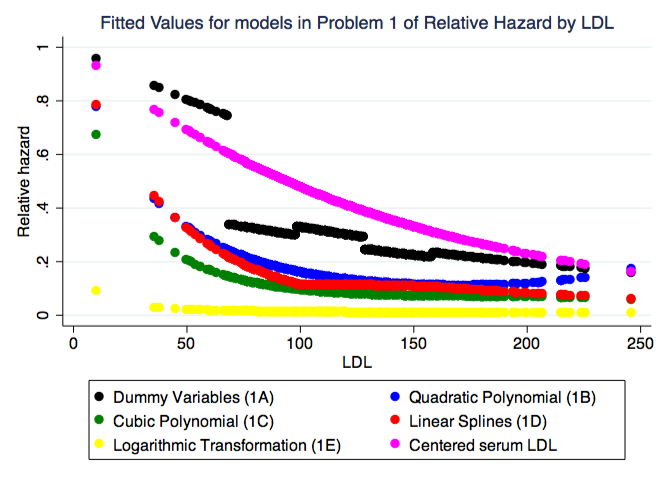
* 1. Fit a regression model in which you test for a linear relationship using linear splines as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save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

I fit a cox proportional hazard model using linear splines with five groups, using 70, 100, 130 and 160 mg/dL as breakpoints for the LDL measurements. With a two-sided p-value of 0.1191, we cannot reject the null hypothesis that all coefficients were equal (i.e., there was no departures from linearity), thus we cannot conclude that the relationship between serum LDL and log hazard function is non-linear.

* 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).

I fit a cox proportional hazard model simultaneously to continuous ldl (ldl – 1) and a logarithmic transformation as an alternative model (the log of LDL). In a testing the hypothesis of a coefficient of zero for the log term, we achieved statistical significance. With a two-sided p-value of 0.004, we can reject the null hypothesis that there were no departures from linearity, and thus can conclude that the relationship is non-linear when 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?



All the models are similar in that they have a general trend in decreasing relative hazard of death associated with an increase in serum LDL. Models A, B, C, and D are pretty similar to each other at high values of LDL (70 – 250 mg/dL) but model A diverges from the rest at lower values of LDL (0-70 mg/dL). Model E predicts values of relative hazard much lower than the other models. The centered serum LDL model is similar to the other models at high LDL values but diverges at mid-level and lower levels of serum LDL. In models 1B, 1C, and 1E, we concluded a non-linear relationship between LDL and relative hazard. In the other models (1E and 1D), we could not conclude nonlinearity.

In this graph, the linear splines model is the most flexible. It approaches the Lowell smooth, and it the most flexible model we performed. The polynomial can include a linear model because it’s a broad model, but it would be best to look at the most flexible model to investigate whether associations are not well described by a linear relationship. Because it’s not forced into a line, we can see how the relationship might change over different LDL ranges.

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.

|  |  |  |
| --- | --- | --- |
| LDL Level | Hazard Ratio | 95% Confidence Interval, P-value |
| 0-70 mg/dL | 0.978 | (0.960 - 0.996, p = 0.019) |
| 70-100 | 0.979 | (0.953 – 1.01, p = 0.131) |
| 100-130 | 0.999 | (0.978 – 1.02, p = 0.934) |
| 130-160 | 0.998 | (0.974 – 1.02, p = 0.875) |
| 160-400 | 0.993 | (0.966 – 1.02, p = 0.678) |

The model exploring the associations between time to death and serum LDL using linear splines estimates the association between time to death and serum LDL between certain defined intervals based on LDL level (in this model, the interval cut off points were defined at 70, 100, 130, and 160 mg/dL). A hazard ratio of less than 1 within each interval indicates a decrease in instantaneous risk of death with an increase of LDL within that interval, and a hazard ratio of greater than 1 within each interval indicates an increase in instantaneous risk of death with an increase of LDL within that interval.

Amongst patients with serum LDL levels between 0 and 70 mg/dL, we see a 2.2% decrease in risk of death associated with a 1 unit increase in LDL. It would not be surprising if the true decrease in risk fell between 4% and 0.04%. With a p-value of 0.019, we reject the null hypothesis of no difference in risk of death associated with an increase in serum LDL between 0 and 70 mg/dL LDL.

Amongst patients with serum LDL levels between 70 and 100 mg/dL, we see a 2.1% decrease in risk of death associated with a 1 unit increase in LDL. It would not be surprising if the true difference in risk fell between a 1% increase and a 4.7% decrease. With a p-value of 0.131, we cannot reject the null hypothesis of no difference in risk of death associated with an increase in serum LDL between 70 and 100 mg/dL LDL.

Amongst patients with serum LDL levels between 100 and 130 mg/dL, we see a 0.1% decrease in risk of death associated with a 1 unit increase in LDL. It would not be surprising if the true difference in risk fell between a 2% increase and a 2.2% decrease. With a p-value of 0.934, we cannot reject the null hypothesis of no difference in risk of death associated with an increase in serum LDL between 100 and 130 mg/dL LDL.

Amongst patients with serum LDL levels between 130 and 160 mg/dL, we see a 0.2% decrease in risk of death associated with a 1 unit increase in LDL. It would not be surprising if the true difference in risk fell between a 2% increase and a 2.6% decrease. With a p-value of 0.875, we cannot reject the null hypothesis of no difference in risk of death associated with an increase in serum LDL between 130 and 160 mg/dL LDL.

Amongst patients with serum LDL levels between 160 and 400 mg/dL, we see a 0.7% decrease in risk of death associated with a 1 unit increase in LDL. It would not be surprising if the true difference in risk fell between a 2% increase and a 3.4% decrease. With a p-value of 0.678, we cannot reject the null hypothesis of no difference in risk of death associated with an increase in serum LDL between 160 and 400 mg/dL LDL.

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

No, there is not evidence that the association between time to death and serum LDL is truly U-shaped. In order for that to be true, we would have to see a negative slope (defined here as a hazard ratio of less than 1, representing a decrease in outcome per increase in predictor variable, as is described in part a of this question) at low LDL levels and a positive slope (defined here as a hazard ratio of greater than 1, representing an increase in outcome per increase in predictor variable, as is described in part a of this question) at high LDL levels. Our results show a statistically significant negative slope (HR = 0.978, p = 0.019) at low LDL levels and a non-statistically significant negative slope (HR = 0.993, p = 0.678) at high LDL levels. It’s possible that there is a positive slope in the higher LDL levels but we could not detect this in our model.

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).

I fit a cox proportional hazards regression modeling the log of serum LDL, race (using dummy variables), and the product of log LDL and race. The instantaneous risk of death (hazard ratio) associated with log LDL for study participants of one race was 0.927, indicating a 7.35% decrease in instantaneous risk of death associated with a doubling of serum LDL while holding race constant. At a constant serum LDL level of 1 mg/dL, being black as opposed to being white (the referent category) was associated with a 84.6% fold decrease in instantaneous risk of death (HR = 0.154), being Asian was associated with a 305.0 fold increase in instantaneous risk of death, and being an other race was associated with a 3.33 x 10^8 fold increase in instantaneous risk of death. A ratio of ratios of 1.36 for black participants indicates a 36% increase in risk of death associated with being black (versus white), *per* doubling of serum LDL. A ratio of ratios of 0.444 for Asian participants indicates a 55.6% decrease in risk of death associated with being Asian (versus white) per doubling of serum LDL level. A ratio of ratios of 0.0615 for participants of other races (versus white) indicates a 93.9% decrease in risk of death associated with being another race, per doubling of serum LDL level.

* 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.

In order to test the hypothesis that race does not modify the association between time to death and serum LDL, after performing the cox proportional hazards regression above, I ran a test comparing if the slopes for any interaction terms were simultaneously zero (the coefficient for black race and LDL interaction, Asian race and LDL interaction, and other race and log LDL interaction). With a two-sided p-value of 0.0009, we reject the null hypothesis that the interaction terms are simultaneously zero (i.e., no effect modification) and can conclude that race does 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.

In order to test the hypothesis that there is no association between time to death a serum LDL, a test of log LDL was run after performing the proportional hazards regression in part a to see if the coefficients for log LDL and any interaction terms that included log LDL were simultaneously zero. With a two-sided p value of less than 0.0001, we reject the null hypothesis of no association between time to death and serum LDL, and can conclude that there is in fact an association between the two.

* 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.

In order to test the hypothesis that there is no association between time to death and race, a test of log LDL was run after performing the proportional hazards regression in part a to see if the coefficient associated with a change in risk of death was associated with different races (modeled as dummy variables with white as the referent race). The test was testing the null hypothesis of the coefficients for any term including race (each dummy variable race term and the interaction terms) were simultaneously zero. With a two-sided p value of less than 0.0001, we reject the null hypothesis of no association between time to death and serum LDL, and can conclude that there is in fact an association between the two.

* 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.

In order to test the hypothesis that there is no difference in the distribution of time to death between whites and blacks, I ran a test testing for a simultaneous zero coefficients for any terms including black race (as a dummy variable, relative to being white, and within the interaction term with log LDL) after running the proportional hazards regression in part a. With a two-sided p-value of 0.5416, we do not have statistical significance, and cannot reject the null hypothesis of 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.

In order to determine sex discrimination in mean salary given in recent years, I tested for the presence of modification of the effect of year on salary by sex. I ran a linear regression of salary including sex, calendar year, and an interaction term between sex and calendar year, modeling sex as a binary and calendar year as a continuous variable. I controlled for correlation within each subject by using the cluster(id) command, and allowed for heterscedasticity by using robust standard errors. To start, on average, women made $675.68 less than men. Among men, the average increase in salary per year was $173.88. On average, per year, there was an increase of $2.85 in female salary relative to male salary. To model the effect of sex, a test was run on the coefficients of the sex term and the sex/year interaction term. With a two-sided p-value of 0.9600, there was no statistical significance, and thus we cannot reject the null hypothesis of simultaneous zero coefficients (i.e., that there is no effect modification). Thus, we cannot conclude that sex modifies the effect of calendar year on faculty salary. To determine the effect of calendar year, a test was run on the coefficients of the year term and the sex/year interaction term. With a two-sided p-value less than 0.0001, we can reject the null hypothesis that within each year, sex has an effect on mean faculty salary. From this, we can conclude that men are always paid more than women, although their salaries increase at similar absolute rates throughout time.

* 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.

In order to determine sex discrimination in geometric mean salary given in recent years, I ran a linear regression modeling the log of salary and sex, year (as a continuous variable), and the interaction between sex and year. I controlled for correlation within each subject by using the cluster(id) command, and allowed for heterscedasticity by using robust standard errors. On average, to start off with, women made 38% less than men. Among men, the average salary increase per 1 year increase was 3.8% of their current salary. There was a 0.40% increase in salary per year for women relative to men. To test the effect of sex on salary over time, I ran a test of the coefficients of the sex term and the term for the sex/year interaction. With a two-sided p-value of 0.7516, we cannot reject the null hypothesis that sex does not modify the effect of calendar year on salary (i.e., sex does not modify the association between time and salary, men and women do not receive differential raises as percents over time. To test the effect of time on salary within each sex, I ran a test of the coefficients of the year term and the term for the sex/year interaction. With a two-sided p-value of less than 0.0001, we can reject the null hypothesis of simultaneous zero coefficients (that year has no effect) and conclude that year does modify the relationship between sex and salary. From this model, we can conclude that women start off making less money than men, and are always making less money than men, although their raises aren’t very different when modeled as a percent.

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

The relative merit of using the model in part a versus b is that we are able to compare differences in absolute raises between men and women over time. The relative merit of using the model in part b versus a is that would be would able to detect differences in percent raises for men and women. We might be interested in assessing each of these, depending on our scientific question.

* 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?

When I ran the models from a and b without controlling for correlation within subject ID, we get the same results, but larger p-values and larger confidence intervals for both. When testing if sex modifies the association between salary and year, the p-values that I got when I didn’t control for correlation within subjects were slightly higher (p = 0.9689, versus 0.9600 for part a, and p = 0.7845, versus 0.7516 for part b), and thus we would be less likely to reject a null hypothesis, making both of these conservative when making inference about associations with sex. For looking at associations by year, both p-values in these models were still < 0.0001, and thus we would still come to the same conclusion and still reject the null hypothesis of no modification of sex on salary by year.