**BIOST 518/515: Applied Biostatistics II/Biostatistics II**

Emerson, Winter 2015

**Homework 06**

March 7, 2015

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:

a. 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).

To test for a linear relationship using a step function as an alternative model, I fitted a Cox proportional hazard model with robust standard errors which included LDL as a continuous term and LDL as dummy variables. The dummy variables included were indicator variables of the following groups: [70, 100), [100, 130), [130, 160) and [160, max LDL]. The reference group for the dummy variables was LDL below 70. I then use the multiple partial F-test to test for departure from linearity, i.e. testing whether the inclusion of dummy variables were significant.

The two-sided p-value from the test was 0.36. Thus we failed to reject the null hypothesis that there was no departure from linearity.

b. 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).

To test for a linear relationship using a quadratic polynomial as an alternative model, I fitted a Cox proportional hazard model with robust standard errors which included LDL as a continuous term and LDL as a quadratic term. The partial Z-test (reported in the R output) was used to test for the departure of from linearity, i.e. testing whether the inclusion of the quadratic term were significant.

The two-sided p-value from the test was 0.054, which is insignificant at 0.05 level. Thus we failed to reject the null hypothesis that there was no departure from linearity.

c. 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).

To test for a linear relationship using a cubic polynomial as an alternative model, I fitted a Cox proportional hazard model with robust standard errors which included LDL as a continuous term, LDL as a quadratic term, and LDL as a cubic term. I then use the multiple partial F-test to test for departure from linearity, i.e. testing whether the inclusion of the quadratic term and the cubic term were significant.

The two-sided p-value from the test was 0.017. Thus we had enough evidence to reject the null hypothesis and concluded that the cubic polynomial would be a better fit to model this data

d. 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).

To test for a linear relationship using linear splines as an alternative model, I fitted a Cox proportional hazard model with robust standard errors which included LDL as a continuous term and LDL as multiple splines. The knots were fixed at 0 (the first knot), 70, 100, 130 and 160. When including all of these terms in the model with LDL as a continuous term, R removed the last spline as thus the fitted value in part f was not based on the last spline. I then use the multiple partial F-test to test for departure from linearity, i.e. testing whether the inclusion of the splines were significant.

The two-sided p-value from the test was 0.12. Thus we failed to reject the null hypothesis that there was no departure from linearity.

e. 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).

To test for a linear relationship using linear splines as an alternative model, I fitted a Cox proportional hazard model with robust standard errors which included LDL as a continuous term and LDL as a natural logarithmic transformed term. The partial Z-test (reported in the R output) was used to test for the departure of from linearity, i.e. testing whether the inclusion of the log-transformed term were significant.

The two-sided p-value from the test was 0.0036. Thus we had enough evidence to reject the null hypothesis and concluded that the log-transformed model would be a better fit to model this data.

f. 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?

**Figure 1. Fitted Relative Hazard Ratios (Reference Group: LDL = 1 mg/dL)**



Except for the models using quadratic polynomial and linear splines, the general tendency of the fitted relative hazard ratios were decreasing as LDL level increasing. In the model using quadratic polynomial, the hazard ratios started rising around LDL level of 180, while in the model using linear splines had a slight upward slope in between LDL of 100 and 150 then decreasing. The quadratic polynomial, cubic polynomial and linear splines model gave pretty close fitted relative hazard ratios.

Based on the fitted values and also on the fact that only the models with cubic polynomial and log-transformed LDL showed departure from linearity, I would prefer the model with only log-transformed LDL. Furthermore, we know from prior belief that the effect of biomedical predictors is typically multiplicative.

2. Consider again a model exploring the associations between time to death and serum LDL when using linear splines.

a. Explain the interpretation of the regression parameters in such a model.

To explore the associations between time to death and serum LDL when using linear splines, the Cox proportional hazard model with robust standard errors was fitted using knots at 70, 100, 130, and 160. The following table presented the results of the estimated coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Response:** **log-hazard** | **Coefficients** | **Exponentiated Coefficients** | **Confident Interval** | **p-value** |
| **Spline 0** | -0.02203 | 0.978 | 0.960 – 0.996 | 0.019 |
| **Spline 70** | -0.02092 | 0.979 | 0.953 – 1.006 | 0.131 |
| **Spline 100** | -0.00091 | 0.999 | 0.978 – 1.021 | 0.934 |
| **Spline 130** | -0.00194 | 0.998 | 0.974 – 1.022 | 0.875 |
| **Spline 160** | -0.00613 | 0.994 | 0.966 – 1.023 | 0.678 |

The coefficient on spline 0 was -0.02203, which implies that among the groups with LDL level between 0 and 70 mg/dL, the hazard ratio between the two groups that differ by 1 mg/dL was 0.978, with group with higher LDL level having lower hazard.

The coefficient on spline 70 was -0.02092, which implies that among the groups with LDL level between 70 and 100 mg/dL, the hazard ratio between the two groups that differ by 1 mg/dL was 0.979, with group with higher LDL level having lower hazard.

The coefficient on spline 100 was -0.00091, which implies that among the groups with LDL level between 100 and 130 mg/dL, the hazard ratio between the two groups that differ by 1 mg/dL was 0.999, with group with higher LDL level having lower hazard.

The coefficient on spline 130 was -0.00914, which implies that among the groups with LDL level between 130 and 160 mg/dL, the hazard ratio between the two groups that differ by 1 mg/dL was 0.998, with group with higher LDL level having lower hazard.

The coefficient on spline 160 was -0.00613, which implies that among the groups with LDL level greater than 160 mg/dL, the hazard ratio between the two groups that differ by 1 mg/dL was 0.994, with group with higher LDL level having lower hazard.

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

There was no evidence that the association between time to death and serum LDL is truly U-shaped. The coefficient on the first spline (spline 0) was statistically significant at 0.05 level (two-sided p-value of 0.019), however, the coefficient on the last spline (spline 160) was not statistically significant (two-sided p-value of 0.678). Also the two coefficient were both negative and not very much different from each other.

3. 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.

a. 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).

A Cox proportional hazard model with robust standard errors was fitted to explore the association between time to death, log-transformed serum LDL, race and their interactions.

The estimated coefficient on log-transformed LDL was -0.0774, which implies that the estimated hazard ratio between two group differed by 1 unit in natural log-transformed LDL (2.72-fold) and being white was 0.461, with group with higher LDL level having lower hazard.

The estimated coefficient on dummy variable for being black was -1.87, which implies that the estimated hazard ratio between the group of black individuals and the group of white individuals with the same LDL level was 0.154, with group of black individuals with similar LDL level having lower hazard.

The estimated coefficient on dummy variable for being Asian was 5.72, which implies that the estimated hazard ratio between the group of Asian individuals and the group of white individuals with the same LDL level was 305.

The estimated coefficient on dummy variable for being other race (not black, white or Asian) was 19.6, which implies that the estimated hazard ratio between the group of other-race individuals and the group of white individuals with the same LDL level was 3.33\*108.

The estimated coefficient on the interaction term between log-transformed LDL and dummy variable for being black was -0.440. This value implies that the estimated hazard ratio between the group of black individuals and having 2.72-fold higher in LDL level and the group of white individuals was 1.55.

The estimated coefficient on the interaction term between log-transformed LDL and dummy variable for being Asian was -1.17. This value implies that the estimated hazard ratio between the group of Asian individuals and having 2.72-fold higher in LDL level and the group of white individuals was 0.310.

The estimated coefficient on the interaction term between log-transformed LDL and dummy variable for being other-race was -4.02. This value implies that the estimated hazard ratio between the group of other-race individuals and having 2.72-fold higher in LDL level and the group of white individuals was 0.0179.

b. 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 for the hypothesis that race does not modify the association between time to death and serum LDL, I performed the multiple partial F-test on the three interaction terms. The two-sided p-value from this test was 0.046, which was statistically significant at 0.05 level. Thus we rejected the null hypothesis and concluded that race modified the association between time to death and log-transformed serum LDL.

c. 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 for the hypothesis that there was no association between time to death and serum LDL, I performed the multiple partial F-test on the log-transformed LDL and the three interaction terms. The two-sided p-value from this test was 0.000026, which was statistically significant at 0.05 level. Thus we rejected the null hypothesis and concluded that there was an association between time to death and serum LDL.

d. 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 for the hypothesis that there was no association between time to death and race, I performed the multiple partial F-test on the three dummy variables for race and the three interaction terms. The two-sided p-value from this test was 0.00000027, which was statistically significant at 0.05 level. Thus we rejected the null hypothesis and concluded that there was an association between time to death and race.

e. 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 for the hypothesis that there was no difference in the distribution of time to death between whites and blacks, I performed the multiple partial F-test on the dummy variable for being black and the interaction term of log-transformed LDL and dummy for being black. The two-sided p-value from this test was 0.54, which is not statistically significant at 0.05 level. Thus we failed to reject the null hypothesis and concluded that there was no difference in the distribution of time to death between whites and blacks.

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

4. 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.

a. 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.

To examine the patterns of raises given to faculty, I used a linear model in which mean salary was modeled by sex, calendar year, and an interaction between sex and calendar year, with adjustment for correlation between data coming from the same individuals. I then used a multiple partial F-test to test for the significance of sex and the interaction between sex and calendar year. The two-side p-value is 2.7\*10-6, which is statistically significant at 0.05 level. Thus we reject the null hypothesis and conclude that there is evidence of sex discrimination in the mean salary given in the recent year.

b. 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.

To examine the patterns of raises given to faculty, I used a linear model in which the geometric mean salary was modeled by sex, calendar year, and an interaction between sex and calendar year, with adjustment for correlation between data coming from the same individuals. I then used a multiple partial F-test to test for the significance of sex and the interaction between sex and calendar year. The two-side p-value is 6.5\*10-8, which is statistically significant at 0.05 level. Thus we reject the null hypothesis and conclude that there is evidence of sex discrimination in the geometric mean salary given in the recent year.

However, the individual p-values of the Z-test on sex and interaction between sex and calendar years were both statistically significant at 0.05 level.

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

The relative merits of the two models:

+ The models allow the possibility of different effect of sex on mean/geometric mean salary in each calendar year (which is realistic in reality, given that salary is affected by the economic condition, which can be represented by calendar year)

+ Model (a) gives an easy-to-understand interpretation of the result and the possibility to derive the total dollar amount of difference in male and female salaries

+ As we expect multiplicative changes in salaries a priori, model (b) will tend to give us more precision, especially given the apparent heteroscedasticy of the salary data.

d. 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?

Had I incorrectly treated the data as independent, the results would still have been significant, however, with greater standard errors, higher p-values and wider confidence intervals on sex, calendar year, and the interaction between sex and calendar year (anti-conservative). The reason is that over time, salary of each individual is expected to be increasing, and therefore, the salary measurement for each individuals will be positively correlated over time.