**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2015

**Homework #6**

March 4, 2015

* 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).

LDL categories were defined by breakpoints at 0, 70, 130 and 160mg/dL. These categories were modeled with centered LDL at 1mg/dL to time to death using proportion hazard regression model with robust standard errors.

We tested the null hypothesis if the model with all the parameters i.e., categorical form of LDL and centered LDL, adequately describe a linear association between log hazard ratio and different forms of LDL. The overall Wald test's two-sided p value was 0.0073 and thus not able to reject the null hypothesis. There is also no evidence for a statistically significant association between LDL levels and time to death using log hazard function.

We used Chi-squared test for linearity to test if all the categories of LDL are equal to zero. The two-sided p value is 0.3609 and thus unable to reject the null hypothesis that the categorical term adequately describes a relationship in which the log hazard function is linear to time to death.

* 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).

Quadratic form of LDL was modeled with centered LDL at 1mg/dL to time to death using proportion hazard regression model with robust standard errors. The two sided p value for overall Chi squared test is 0.0005, showing strong evidence of a statistically significant association between LDL and time to death using log hazard function.

We used Chi-squared test for linearity to test this second order form of LDL is equal to zero. The two-sided p value is 0.055, and thus unable to reject that this term adequately describes a relationship in which the log hazard function is linear to time to death.

* 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).

Cubic form of LDL was modeled with centered LDL at 1mg/dL to time to death using proportion hazard regression model with robust standard errors. The two sided p value for overall Chi squared test is <0.0001, showing strong evidence of a statistically significant association between LDL and time to death using log hazard function.

We used Chi-squared test for linearity to test if the second and third order form of LDL are equal to zero. The two-sided p value is 0.0164, thus rejecting the presence of linear association with time to death. The cubic form of LDL is adequately modeled by a relationship in which the log hazard function is not linear to time to death.

* 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).

Spline forms of LDL with knots at 0,70,100,130,160,400mg/dL were modeled with centered LDL at 1mg/dL to time to death using proportion hazard regression model with robust standard errors. The two sided p value for overall Chi squared test is <0.0001, showing strong evidence of a statistically significant association between LDL and time to death using log hazard function.

We used Chi-squared test for linearity to test if the spline forms of LDL are all equal to zero. The two-sided p value is 0.1191, thus unable to reject the presence of linear association with time to death. The spline form of LDL is adequately modeled by a relationship in which the log hazard function is linear to time to death.

* 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).

Log transformed LDL to the natural base was modeled with centered LDL at 1mg/dL to time to death using proportion hazard regression model with robust standard errors. The two sided p value for overall Chi squared test is <0.0001, showing strong evidence of a statistically significant association between LDL and time to death using log hazard function.

We used Chi-squared test for linearity to test if the log transformed LDL is equal to zero. The two-sided p value is 0.0036, thus rejecting the null hypothesis of presence of linear trend in association between hazard ratio of death and LDL. The log-transformed form of LDL is adequately modeled by a relationship in which the log hazard function is not linear to time to death.

* 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?

The models with quadratic, cubic and spline forms of LDL are similar, with a curvilinear association and quadratic with almost u-shaped association. The quadratic form has a more smooth form compared to spline. The step function model does not seem to fit the data while logarithmic form shows no association.

The choice of a particular model could be based on the smoothness of the curves; the step and splines do not seem to have smooth curves and thus excluded from choice. In addition, the log form shows no association. The overall chi square test in the model exclude centered LDL form, leaving the graphs with the quadratic and cubic form describe a linear relationship with LDL and time to death using a log hazard function. The choice of model will then be based on parsimony to avoid inflation of alpha 1 error and thus the model with least parameters (quadratic form) will be my choice.



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.

Among those with LDL levels between 1 to 70mg/dL, the group that differs with a higher 1 mg/dL of LDL has a lower death risk of 0.9842 times those with lower LDL levels.

Among those with LDL levels between 70 to 100mg/dL, the group that differs with a higher 1 mg/dL of LDL has a lower death risk of 0.9853 times those with lower LDL levels.

Among those with LDL levels between 100 to 130mg, the group that differs with a higher 1 mg/dL of LDL has a higher death risk of 1.005231 times those with lower LDL levels.

Among those with LDL levels between 130 to 160mg, the group that differs with a higher 1 mg/dL of LDL has a higher death risk of 1.004187 times those with lower LDL levels.

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

There is no evidence of a U-shaped association. The negative slopes of sldlA and sldlB (Hazard Ratio less than 1) and positive slopes (Hazard Ratio greater than 1) of sldlC and sldlD are not significant.

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).

Please note that we used Note that we used log transformation (base 2) for ldl for question 3 (all the sections)

* For each two-fold increase in serum LDL among whites, there is 0.5846 times lower death risk in the group with higher LDL levels.
* The death risk among blacks with LDL level of 1mg/dL is 0.1502 times lower to the death risk of whites with LDL level of 1mg/dL.
* The death risk among Asians with LDL level of 1mg/dL is 304.98 times higher to the death risk of whites with LDL level of 1mg/dL.
* The death risk among Asians with LDL level of 1mg/dL is 3.33 \* 108 times higher to the death risk of whites with LDL level of 1mg/dL.
* 1.356486 is the ratio of two hazard ratios comparing death risk of two groups of blacks differing by two fold increase of LDL levels to death risk of two groups of whites differing by two fold increase in LDL
* 0.4442549 is the ratio of two hazard ratios comparing death risk of two groups of asians differing by two fold increase of LDL levels to death risk of two groups of whites differing by two fold increase in LDL
* 0.0615902 is the ratio of two hazard ratios comparing death risk of two groups of "other" race differing by two fold increase of LDL levels to death risk of two groups of whites differing by two fold increase in LDL

* 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.

We used Chi-squared test to examine if the coefficient of interaction terms of race are all equal to zero. The two-sided p value is 0.0452, rejecting the null hypothesis. This parameter 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.

We used Chi-squared test to examine if log transformed LDL and interactions terms are all equal to zero. The two-sided p value is <0.0001 thus rejecting the null hypothesis. There is an 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.

We used Chi-squared test to examine if the coefficients of interaction terms and race are all equal to zero. The two-sided p value is <0.0001, thus rejecting the null hypothesis. There is an 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.

We used testparm (chi square test) with null hypothesis that there is no difference in the distribution of time to death between whites and blacks. We tested the dummy variable for blacks and interaction term for black and log transformed LDL are equal to zero. The two-sided p value was 0.5416, thus not able to reject the null hypothesis of no difference in distribution of time to death between whites and blacks.

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

* 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.

There is no evidence of mean salary differences between men and women; women have lower salary with $602.65 less dollars (95% CI 1,209.12 lower to 3.82 higher) compared to men in the same calendar year but this finding is not statistically significant (p=0.051).

The difference between the mean salary difference between women differing with one calendar year to the mean salary difference of men differing with one calendar year is $11.79459 ( 95% CI is 122.01 higher to $98.42 lower salaries between women to men), and is not statistically significant with p value of 0.833.

* 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.

There is evidence of geometric mean salary differences between men and women; there is a 13.50% reduction in geometric mean salary for women compared to men in same calendar year (95% CI $0.76 - $0.99) and is statistically significant with two-sided p value of 0.032.

Men have 3.66% higher salaries with one year increment (95% CI 1.71% - 5.70%), this is statistically significant with a two sided p value <0.0001.

The ratio of geometric mean ratios between women differing with one calendar year to men differing with one calendar year is 0.68% higher for women compared to men. This is however not statistically significant (p=0.576).

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

Model a that uses mean salary difference will be useful to determine the magnitude of the gap between salaries in dollar amount that is could be easily used to make changes in polices and compensation packages. In addition, the increase in dollar amount is more meaningful because the salary amount for women to start with is low and geometric increases of even 10% can give a false sense that women have higher increases compared to men and obscures the lower actual amount received.

The merits of Model b that uses geometric mean as a difference also can provide meaningful differences; an "x-fold" difference between salaries may have a stronger impact and memorable figure to policy makers. In addition, salary increments in the real word are not in absolute dollar amount but a percentage of salary and this model can capture this "natural" increments per year for each sex.

* 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 had treated the observations as independent without accounting for correlation, the difference of mean salaries between women and men in same calendar year would be significant unlike model that accounted for correlation.

The models (parameters labeled by a star) that do not account for correlation tended to have smaller SE, narrower 95% CI and anticonservative p values as depicted in tables below.

\* model treating data as independent

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model A | coefficient | SE | t | p | 95% CI | |  |
| female | -602.6514 | 306.2281 | -1.97 | 0.051 | -1209.12 | 3.817335 |  |
| female\* | -602.6514 | 250.2724 | -2.41 | 0.016 | -1094.493 | -110.8096 |  |
| cyear | 169.2224 | 46.39253 | 3.65 | 0.000 | 77.34443 | 261.1004 |  |
| cyear\* | 169.2224 | 54.95619 | 3.08 | 0.002 | 61.22107 | 277.2237 |  |
| interaction | 11.79459 | 55.65021 | 0.21 | 0.833 | -98.41774 | 122.0069 |  |
| interaction\* | 11.79459 | 71.09067 | 0.17 | 0.868 | -127.9147 | 151.5038 |  |
| constant | 4422.936 | 245.7341 | 18.00 | 0.000 | 3936.272 | 4909.599 |  |
| constant\* | 4422.936 | 195.91 | 22.58 | 0.000 | 4037.928 | 4807.943 |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model B | coefficient | SE | t | p | 95% CI | |  |
| female | .8650472 | .0578415 | -2.17 | 0.032 | .7577558 | .9875302 |  |
| female\* | .8650472 | .0444875 | -2.82 | 0.005 | .7818921 | .957046 |  |
| cyear | 1.036648 | .0101635 | 3.67 | 0.000 | 1.016714 | 1.056973 |  |
| cyear\* | 1.036648 | .0112768 | 3.31 | 0.001 | 1.014722 | 1.059048 |  |
| interaction | 1.006755 | .0121011 | 0.56 | 0.576 | .9830726 | 1.031008 |  |
| interaction\* | 1.006755 | .0141743 | 0.48 | 0.633 | .9792813 | 1.035 |  |
| constant | 4301.373 | 229.021 | 157.14 | 0.000 | 3870.904 | 4779.714 |  |
| constant\* | 4301.373 | 173.1407 | 207.86 | 0.000 | 3974.223 | 4655.454 |  |