Biostatistics 518

Homework #6

6079

1. Serum LDL and time to death.
   1. Test for linear relationship using a step function as an alternative model.
      1. I performed proportional hazards regression with robust standard errors for time to death by serum ldl including ldl modeled as both a continuous variable (with reference group of 1) and modeled as a dummy variable with categories of 0-69, 70-99, 100-129, 130-159, 160-399, 400+.
      2. Given a p-value of 0.3609 from the multiple partial F test we cannot reject the null hypothesis that the difference from the linear regression is zero in all groups defined by the dummy variables. (cannot reject the null hypothesis the the departure from linearity in all groups =0).
   2. Test for linear relationship using a quadratic polynomial as an alternative model.
      1. I performed proportional hazards regression with robust standard errors for time to death by serum ldl including ldl modeled as both a continuous variable (with reference group of 1) and modeled as a quadratic polynomial (ldl-1)2.
      2. Given a p-value of 0.055 we cannot reject the null hypothesis that there is a no higher order relationship between the squared value of ldl and time to death when adjusting for the linear relationship.
   3. Test for linear relationship using a cubic polynomial as an alternative model.
      1. I performed proportional hazards regression with robust standard errors for time to death by serum ldl including ldl modeled as a continuous variable (with reference group of 1), a and a cubic polynomial (including both(ldl-1)2 and (ldl-1)3).
      2. Given a p-value from the multiple partial F-test testing ldl2 and ldl3 of 0.0164 we can reject the null hypothesis that there is no higher order (ldl squared or ldl cubed) association between ldl and time to death when adjusting for the linear relationship.
   4. Test for linear relationship using linear splines as an alternative model.
      1. I performed proportional hazards regression with robust standard errors for time to death by serum ldl including ldl modeled as linear splines with knots at 70, 100, 130, and 160.
      2. The chi2 test for equality of slopes in each group gives a p-value of 0.1191, therefore we cannot reject the null hypothesis that all slopes are equal (that the association is linear).
   5. Test for linear relationship using a log-transformed ldl as the alternative model.
      1. I performed proportional hazards regression with robust standard errors for time to death by serum ldl including ldl modeled as a continuous variable (with reference group of 1), and log(e) transformed ldl.
      2. Given a p-value of 0.004 we can reject the null hypothesis that there is a definite linear relationship between ldl and time to death in favor of the alternative that the relationship may be better described by a relationship between time to death and loge(ldl).
   6. Fitted values for models a-e above and centered ldl.



* + 1. First order trend of decrease in hazard ratio with increase in LDL is similar, but the fitted values are different depending on what model you use. The hazard ratios for the log(ldl) model are the lowest overall, followed by the cubic model, the linear splines, and quadratic model, the dummy variables, and then the centered LDL. The centered LDL predicts a much higher hazard for each value of LDL, likely because it predicts VERY low and even negative values that are not found in our data set or in nature.
    2. The second order trend shows a curvature such that the increase in hazard ratio with decreasing LDL begins to become non-linear at an LDL of around 100. Only the dummy variable groups is flexible enough to show the much higher hazard in the LDL group < 70.
    3. If I had to choose one of these analyses *a priori* I think I would choose the model using linear splines because it is flexible enough to show a linear relationship if one exists, but still allows for non-linearity.

1. A model exploring the associations between time to death and serum LDL using linear splines.
   1. Interpretation of the regression parameters:
      1. β0 (the intercept), eβ0 = the instantaneous risk of death when X=0 (in this case when LDL=0)
      2. Slope parameters: Βi gives the log hazard ratio for measurements of LDL between the same 2 knots but differing by 1 unit in X (ldl). eβi = estimated hazard ratio for measurements of LDL between the same 2 knots but differing by 1 unit of LDL
   2. Is there evidence that the association between time to death and LDL is U-shaped?
      1. In the groups defined by each knot, all hazard ratios are less than 1. However, the only p-value of significance is the p-value (0.019) for spline group 1 (0-69). If there were a U-shaped association, we would expect the hazard ratio to be in different directions at the extremes of the ldl values (<1 for lower ldl and > 1 for higher ldl or vice-versa) and we would also need to show p-values <0.05 for those hazard ratios. Based on this model, we have no evidence that there is a U-shaped association between serum ldl and time to death. However, we also have no evidence that there is not a U-shaped association.
2. Exploring the associations between time to death and serum LDL as possibly modified by race.
   1. Model of time to death regressed on log(2)-transformed LDL, race, and their interaction. Provide explicit interpretation of each parameter in your model.
      1. eβ1 (coefficient for log2ldl) = 0.585 corresponds to the hazard rate for a doubling in LDL when race=reference group (white).
      2. eβ2 (coefficient for race=2) = 0.154 corresponds to the hazard ratio for race=2 (blacks) compared to race=1 (whites) when LDL=1.
      3. eβ3 (coefficient for race=3) = 304.98 corresponds to the hazard ratio for race=3 (Asians) compared to race=1 (whites) when LDL=1
      4. eβ4 (coefficient for race=4) = 3.3 X 108 corresponds to the hazard ratio for race=4 (others) compared to race=1 (whites) when LDL=1.
      5. Coefficients for 3 interaction terms: These are all ratios of ratios since we are performing proportional hazards regression.
         1. Race#c.logldl2: The ratio of the hazard ratio among black people with a 1 unit difference in log2LDL to the hazard ratio among white people with a 1 unit difference in log2LDL is 1.55.
         2. Race#c.logldl3: The ratio of hazard ratio among asian people with a 1 unit difference in log2LDL to the hazard ratio among white people associated with a 1 unit difference in log2(LDL) is 0.31.
         3. Race#c.logldl4: The ratio of hazard ratio among people of “other” races with a 1 unit difference in log2LDL to the hazard ratio among white people associated with a 1 unit difference in log2(LDL) is 0.018
   2. Use 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:

I performed the post-regression F-test using only the interaction term i.race#c.logldl. This returned a 2-sided p-value of 0.0452 which indicates that we can reject the null hypothesis that the coefficient for each of the interaction terms is zero. There may be effect modification by race in the association between log2ldl and hazard of death.

* 1. Use 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:

I performed the post-regression F-test using log2ldl and the interaction term race\*log2ldl. With a 2-sided p-value of <0.0001, we can reject the null hypothesis that there is no association between log2ldl and time to death.

* 1. Use the regression analysis in part a to perform a statistical test of the hypothesis that there is not association between time to death and race:

I performed the post-regression F-test using race and the interaction term race\*log2ldl. With a 2-sided p-value of <0.0001 we can reject the null hypothesis that there is 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 not difference in the distribution of time to death between whites and blacks.

I performed a post-regression F-test using black race only (given that white is the reference group). With a p-value of 0.542, we cannot reject the null hypothesis that there is no difference in the distribution of time to death among blacks when compared to whites.

1. University salary dataset questions
   1. Is there evidence of sex discrimination in the mean salary given in recent years?
      1. I re-parameterized the data such that 1990 was represented by 0, 1991 by 1, 1992 by 2, etcetera, such that the reference group was 1990. I performed linear regression for the association between sex and mean salary adjusted for year and sex\*year interaction term. I accounted for correlated data by clustering by study id number.
      2. The intercept, which represented mean salary in men in 1990, was $4368. Women were paid a mean $550 less than men in 1990.
      3. I performed a post-regression test evaluating the null hypothesis that the difference of: the difference in mean salary in women compared to men in one year and the difference in mean salary between women and men in the previous year, was zero. The F-test for the interaction term revealed a p-value of 0.9868.
      4. Based on these results, it would appear that there may be discrimination in starting salary by some mechanism, but that the mechanism is not a difference in the rate of change in salary between women and men.
   2. Is there evidence of sex discrimination in the geometric mean salary given in recent years?
      1. The geometric mean salary for men in 1990 was estimated at $4,230 and women were paid 13.4% less than men.
      2. The post-regression test revealed a p-value of 0.6837, indicating that we cannot reject the null hypothesis that the coefficient for the interaction term is zero.
      3. Again, as above, there may be discrimination in starting salary by some mechanism, but there does not appear to be a difference in the rate of change between the sexes.
   3. What are the relative merits of the 2 models used in parts a and b?
      1. The merit of model a is that it is interpretable in real terms. For example, if one wanted to determine the average amount by which women’s salaries would need to be increased (or by what percentage per year to equal men’s salaries in 5 year or 10 year or what have you), it would be easy to calculate and make real changes in the real world.
      2. The merit of model b is that it downweights any salary outliers and therefore may improve the precision of the estimates. Also, a small change in the difference may be seen as a larger change in the ratio of geometric means, making any interaction between sex and salary more obvious.
   4. Because monthly salary will tend to increase over time in the same individual, the data points will be positively correlated. When data is positively correlated and the statistic adds the data (as with a mean) and you do not account for the correlation, your estimates will tend to be anti-conservative. When we run this analysis clustered by id number, we get a SE for the mean salary difference of 275, but when run assuming independence we get SE of 227. Thus, as expected, when we do not presume independence are results are anti-conservative. If we were instead to cluster by year, the impact on the standard error is in the opposite direction (when we do not account for correlated data the estimate is conservative).