Biost 515/518

Homework 6

**Methods for every problem:** It is assumed I used these methods unless otherwise specified in the problem answer. I used a .05 level of significance for all inference. Data was provided on 735 subjects of which 10 subjects were missing data for Ldl levels. These 10 subjects were omitted from all analyses. It should be noted that I coded in R for all of my answers, so they may not match exactly the same as the STATA output. When partitioning Ldl levels, I partitioned as follows: [0 – 70], (70 – 100], (100 – 130], (130 – 160], 161+. That is, when grouping, the boundary Ldl values were included in the lower group (this can make a big difference in p-values).

**1. a) Methods:** I performed Cox Proportional Hazards Regression to test for a linear relationship between the log(hazard) and LDL. My response was all-cause mortality which was regressed on LDL as a continuous variable and LDL as categorical, dummy variables. I used the Wald test with robust standard errors to perform a multiple partial test, testing whether all of the coefficients of the categorical variables were zero.

**Results:** The two-sided p-value for testing whether all of the coefficients of the categorical variables were zero was .567. This does not provide evidence that the relationship between log(hazard) and LDL is non-linear. We do not conclude that the relationship is linear, however. The relationship between log(hazard) and LDL could be non-linear in a way that our model could not determine.

**b) Methods:** I performed Cox Proportional Hazards Regression to test for a linear relationship between the log(hazard) and LDL. My response was all-cause mortality which was regressed on LDL and LDL squared as continuous variables. I used the Wald test with robust standard errors to test whether the coefficient of the squared term was zero.

**Results:** The two-sided p-value for testing whether the coefficient of the squared term was zero was .055. This does not provide evidence that the relationship between log(hazard) and LDL is non-linear (significance level of .05). We do not conclude that the relationship is linear, however. The relationship between log(hazard) and LDL could be non-linear in a way that our model could not determine.

**c) Methods:** I performed Cox Proportional Hazards Regression to test for a linear relationship between the log(hazard) and LDL. My response was all-cause mortality which was regressed on LDL, LDL squared, and LDL cubed as continuous variables. I used the multiple partial Wald test with robust standard errors to test whether the coefficients of the squared term and cubic term were both zero.

**Results:** The two-sided p-value for testing whether the coefficients of the squared term and cubic term were zero was .017. This provides evidence that the relationship between log(hazard) and LDL is non-linear (significance level of .05).

**d) Methods:** I performed Cox Proportional Hazards Regression to test for a linear relationship between the log(hazard) and LDL. My response was all-cause mortality which was regressed on LDL as a continuous variable and using linear splines. I used the multiple partial Wald test with robust standard errors to test whether the coefficients of the spline terms were zero.

**Results:** The two-sided p-value for testing whether the coefficients of the spline terms were zero was .128. This does not provide evidence that the relationship between log(hazard) and LDL is non-linear (significance level of .05). We do not conclude that the relationship is linear, however. The relationship between log(hazard) and LDL could be non-linear in a way that our model could not determine.

**e) Methods:** I performed Cox Proportional Hazards Regression to test for a linear relationship between the log(hazard) and LDL. My response was all-cause mortality which was regressed on LDL as a continuous variable and LDL as a log base e transformed continuous variable. I used the Wald test with robust standard errors to test whether the coefficient of the log transformed variable is zero.

**Results:** The two-sided p-value for testing whether the coefficient of the log transformed variable was zero was .004. This provides evidence that the relationship between log(hazard) and LDL is non-linear (significance level of .05).

**f)** I have fit the various models on the same axis. Here is that graph:



\* Gold = Log transformed \* Purple = Centered Ldl \* Red = Step function

\* Blue = Quadratic \* Green = Cubic

The models which provided evidence against linearity between log transformed hazard and LDL were the cubic model and the log transformed model with much stronger evidence coming from the log transformed model. This suggests that when we do not think a relationship to be linear, log transforming the POI may be the best way to model the relationship. If we do think there is a U-shaped relationship than maybe we will also fit a cubic model. However, the other models do not provide strong evidence one way or another about the linearity of the relationship.

**2. a)** The coefficients of the linear spline terms is the slope of the best fitting line within only that range of data subject to the constraints that all lines must meet on their boundaries. Therefore, for the spline term in our model from 0-70, the coefficient is the slope of the best fitting linear relationship between log hazard and LDL for LDL values between 0 and 70. This allows the slope to change throughout the model while remaining constant for certain ranges of values.

**b)** No, there is no evidence that our association between log hazard and LDL is truly U-shaped. The model which would have detected this best is the cubic model. However, even the cubic model did not fit a U-shaped relationship. The cubic looks to mimic the log transformed shape, and the evidence against linearity was not as strong as the log transformed variable.

**3. a)** I fit a linear regression model regressing hazard on centered LDL and race as a categorical dummy variable. The coefficient of -.007 for the centered LDL term reflects a 0.7% decrease in hazard for each 1 point increase in LDL when race = 1. When race = 2, we add the -.007 coefficient for LDL to the .005 coefficient for the interaction between race = 2 and LDL to get a 0.2% decrease in hazard for every 1 point increase in LDL when race = 2. When race = 3, we add the -.007 coefficient for LDL to the -.011 coefficient for the interaction between race = 3 and LDL to get a 1.8% decrease in hazard for every 1 point increase in LDL when race = 3. When race = 4, we add the -.007 coefficient for LDL to the -.039 coefficient for the interaction between race = 4 and LDL to get a 4.5% decrease in hazard for every 1 point increase in LDL when race = 4.

**b)** I performed a likelihood ratio test with robust standard errors to determine if all interaction terms were simultaneously zero. With a p-value of .182, we cannot reject the hypothesis that all interaction terms are simultaneously zero. Therefore, this does not provide evidence that race modifies the association between all cause death and LDL.

**c)** I performed a likelihood ratio test with robust standard errors to determine if all interaction terms and the continuous LDL term were simultaneously zero. With a p-value of .012, we can reject the hypothesis that all interaction terms and the continuous LDL term are simultaneously zero. Therefore, this does provide evidence that there is an association between all-cause death and LDL.

**d)** I performed a likelihood ratio test with robust standard errors to determine if all interaction terms and all race categorical terms were simultaneously zero. With a p-value of .349, we cannot reject the hypothesis that all interaction terms and race categorical terms are simultaneously zero. Therefore, this does not provide evidence that there is an association between all-cause death and race.

**e)** In my regression, race = white is the reference for the dummy categorical variable. Therefore, I have a term in my model which compares the difference between white and black. I performed a multiple partial F test on this term as well as the term for interaction between the comparison between white and black and LDL. With a p-value of (I was not able to get a p-value for r).

**4. a)** No, there is no evidence of sex discrimination is granting salary increases given to faculty hired in recent years after restricting our sample as required. I based my answer from a linear regression model modeling salary by sex, start year and an interaction between sex and start year adjusting for correlated measurements. I performed a multiple partial F-Test to test if the coefficients for sex and the interaction term were simultaneously zero. My p-value was not significant, which provides no evidence of sex discrimination in granting salary increases. I did adjust for correlated data.

**b)** No, there is no evidence for sex discrimination modeling the geometric mean salary. I performed the analysis as in part **a)** after log transforming salary. I still got a non-significant p-value.

**c)** Modeling the geometric mean down-weights outliers, and models increases in salary on a multiplicative scale. Our p-value for sex discrimination was smaller when modeling the geometric mean, which may indicate that modeling the geometric mean salary has higher power to discover sex discrimination in salary increases. Also, we do think salaries increase on a multiplicative scale, so modeling the geometric mean probably gives us more accurate results.

**d)** Since salary increases are positively correlated, and we are modeling the mean (adding the values), not accounting for correlated data will increase our standard error estimates. This will make our estimates overly conservative. This is also the case when providing inference about the asociation between salary increases and start year.