Homework #6

12 March 2015

**Question 1**

* 1. Proportional hazards regression with robust standard errors was used to model the relationship between LDL and time to death, modeling LDL as a linear term and as a categorical term with LDL split at 70, 100, 130, 160, and 400. A two-sided p-value of 0.36 for the multiple partial F test testing all of the dummy variables for the categorical LDL variable does not provide evidence of a non-linear relationship.
	2. Proportional hazards regression using robust standard errors was used to model the relationship between LDL and time to death, modeling LDL with both a linear term and quadratic term in the model. The two-sided p-value for the partial z test of the coefficient of the quadratic term (p=0.0550) provides moderate, but not strong (it failed to reach statistical significance), evidence that the relationship that the relationship is non-linear.
	3. Proportional hazards regression with robust standard errors was used to model the relationship between LDL and time to death, modeling LDL with both a linear term and a cubic term. A two-sided p-value for the coefficient of the cubic term (p=0.112) does not provide evidence of a non-linear relationship between LDL and time to death.
	4. Proportional hazards regression with robust standard errors was used to model the relationship between LDL and time to death, modeling LDL with both a linear term and linear splines with knots at 70, 100, 130, and 160 mg/dL. A non-significant two-sided p-value (p=0.369) for the multiple partial F test including all of the splines does not provide evidence of a non-linear association between LDL and time to death.
	5. Proportional hazards regression with robust standard errors was used to model the relationship between LDL and time to death, modeling LDL as both linear term and a log-transformed term. The two-sided p-value for the coefficient of log-transformed term (p=0.004) provides strong evidence of a non-linear relationship between LDL and time to death.
	6. Predicted values from models in parts a through e are shown in the figure below. Each of these models provide much more similar predicted values in the middle of the range of LDL values compared to very high or very low values. Linear splines are similar to lowess smooths. If the most flexible model (for example the cubic polynomial) looks similar to the less flexible model (only using the linear term) then this suggests that the association is adequately modeled by the simpler model and thus there is no reason to use the more complex model.



**Question 2.**

1. In a model using linear splines to model the relationship between serum LDL and time to death, the estimated hazard ratios are as follows. For groups differing in LDL by 1 mg/dL, the estimated hazard ratio is 0.978 when the LDL is between 0 and 70 mg/dL, 0.979 when the LDL is between 70 and 100 mg/dL, 0.999 when the LDL is between 100 and 130 mg/dL, 0.998 when the LDL is between 130 and 160 mg/dL, and 0.994 when the LDL is greater than 160 mg/dL, with the lower instantaneous risk of death in the group with the higher LDL.
2. When using linear splines, evidence of a u-shaped relationship would be a difference in the sign of the slope in categories at either end of the range of LDL values. In this model we see that the estimated hazard ratio is in fact relatively similar between those with an LDL less than 70 (HR=0.978) and those with an LDL greater than 160 (HR=0.994). Based on this model, there is no evidence of a u-shaped relationship between LDL and time to death.

**Question 3.**

1. When comparing two groups of whites with a two-fold difference in serum LDL, we estimate a 20.8% lower instantaneous risk of death in the group with the higher LDL (HR=0.792). When comparing blacks versus white when both groups have a serum LDL of 0 mg/dL, we estimate an 84.6% lower instantaneous risk of death among blacks (HR=0.154). When comparing Asians to whites when both groups have a serum LDL of 0 mg/dL, we estimate that the instantaneous risk of death is 305 times higher among Asians (HR=305). When comparing subjects of ‘other’ races to whites when both groups have a serum LDL of 0 mg/dL, we estimate a hazard ratio of 3.33x10-8.

Here the coefficients for the interaction terms can be interpreted as ratios of ratios. 1.55 is the ratio of the hazard ratio among blacks differing by an e-fold difference in LDL to the hazard ratio among whites differing by an e-fold difference in LDL. 0.310 is the ratio of the hazard ratio among asians differing by an e-fold difference in LDL to the hazard ratio among whites differing by an e-fold difference in LDL. 0.0179 is the ratio of the hazard ratio among those of ‘other’ races differing by an e-fold difference in LDL to the hazard ratio among whites differing by an e-fold difference in LDL.

1. The two-sided p-value for the multiple partial chi-squared (p=0.0452) provides statistically significant evidence that the coefficient for the interaction term is not equal to zero and thus we conclude that race modifies the association between serum LDL and time to death.
2. The two-sided p-value for the multiple partial chi-squared test (p<0.00005) provides strong evidence for rejecting the null hypothesis of no association between serum LDL and time to death.
3. The two-sided p-value for the multiple partial chi-squared test (p=0.0452) provides statistically significant evidence for rejecting the null hypothesis that the coefficients for race as a main effect and as an interaction with LDL are both zero, in favor of the alternative hypothesis of an association between race and time to death.
4. In order to test the null hypothesis of no difference in the distribution of time to death between whites and blacks, we test the coefficients for both the white and black dummy variables and the interaction term between black and white race and LDL. A two-sided p-value of 0.0452 provides evidence for rejecting the null hypothesis in favor of the alternative hypothesis of a difference in the distribution of time to death between whites and blacks.

**Question 4.**

* 1. We used linear regression with robust standard errors and accounting for correlated data to estimate the difference in mean salary between males and females. The model we used included calendar year and the interaction of sex with calendar year in addition to the predictor of interest, sex. From this regression, a two-sided p-value of 0.0687 testing both the sex parameter and the interaction term, we reject the null hypothesis of no association between sex and salary.
	2. We used linear regression with robust standard errors and accounting for correlated data to estimate the difference in geometric mean salary between males and females. The model we used included calendar year and the interaction of sex with calendar year in addition to the predictor of interest, sex. From this regression, a two-sided p-value of 0.0005 provides strong evidence for rejecting the null hypothesis of no difference in geometric mean salary between men and women.
	3. The model in part a using mean as a summary measure is preferable when we are interested in the absolute quantity of money difference between the two groups. Additionally, it is often easier to communicate to a wider audience about differences in means. However, due to the skewed distribution of salary, inference about the geometric mean can be more precise. The geometric mean is also more relevant when interested in proportionate rather than absolute differences.
	4. Had we not accounted for the correlated data in the above models, we would have seen a p-value of p<0.00005 for both parts a and b. Failing to account for correlated data (ie. treating all observations as independent) would have tended to produce anti-conservative inference about associations sex. However, when considering associations by year within only men or only women, we would tend to have more precision with these correlated data.