BIOST 518 Homework 6

1. For all analyses undertaken for this problem, ten subjects out of the 735 subjects were excluded due to missing data.

	1. The linear association between observation time and serum ldl was tested using ldl modeled as a dummy variable as the alternative model. First, a model including both continuous serum ldl and ldl as the step-wise portions from 0-70 , 70-100, 100-130, and 130-160 mg/dL. A second regression was carried out using only the continuous ldl measurements as the predictor. The likelihood ratio test was then performed to compare the two models based on their “fits”. The two-sided p-value obtained for the test is 0.8963.
	2. Linear regression was used to test the relationship between time to death and serum ldl, modeling serum ldl as a quadratic function as the alternative model. A likelihood ratio test was used to compare a model containing both ldl as a continuous and as a quadratic function with a model that only contains a ldl as a continuous variable. The two-sided p value obtained was 0.4062.
	3. Linear regression was used to test the relationship between time to death and serum ldl, modeling serum ldl as a cubic function as the alternative model. A likelihood ratio test was used to compare a model containing both ldl as a continuous and as a cubed function with a model that contained only ldl as a continuous variable. The two-sided p value obtained was 0.3842.
	4. Linear regression was used to test the relationship between time to death and serum ldl, modeling serum ldl as linear splines with knots at 70, 100, 130, and 160 mg/dL as the alternative model. A likelihood ratio test was then used to compare a model containing both ldl as a continuous and as linear splines with a model that contained only ldl as a continuous variable. The two-sided p value obtained was 0.7323.
	5. Linear regression was used to test the relationship between time to death and serum ldl, modeling serum log-transformed ldl as the alternative model. A likelihood ratio test was then used to compare a model containing both ldl as a continuous and as a log-transformed variable with a model that contained only ldl as a continuous variable. The two-sided p value obtained was 0.5862.
	6. Although the fitted values have distinct overlapping sections, each one is quite distinct from the model utilizing the centered ldl. This indicates that simply fitting the data onto a linear model is likely not the best route for data analysis. In looking at the model utilizing linear splines, we can see that the slopes for the estimates differ depending on the LDL interval being looked at. When looking at whether associations are well described by a linear model, having the different models will inform if this is such a good choice. Clear deviations indicating a non-linear relationship indicates that a new model may be needed to describe the relationship, one that is non-linear.

	
	7. The intercept for this type of analysis would still refer to the serum LDL levels when the levels are 0 mg/dL. The obtained coefficients for each spline section, however, are still slopes, however, they only apply to the given interval. For example, in an interval from 0 to 70 mg/dL of serum LDL, the slope obtained for that portion is the mean difference in time to death between groups that differ in serum LDL by 1 mg/dL, but only applies to serum LDL from 0 to 70 mg/dL.
	8. There is some evidence for a U-shaped association between time to death and serum LDL, as indicated by the fitted values when modeling using splines. This also makes some sense scientifically as LDL levels are optimal at a specific range. If they are too low or too high, then health problems may ensue, shortening the time to death.
	9. Based on the regression model fitting both the log (base 2)-transformed serum ldl values and adjusting for race as a dummy variable (using white as the reference group), we find that for each two-fold increase in serum ldl levels, it is expected that the time to death will increase by 77 days with a baseline level of 1281 days until death among individuals with no serum LDL at all. Among blacks, this increase is expected to be 51 days lower than whites, while among Asians the increase is estimated to be 21 days greater than whites. In any other racial group, the effect on time to death is expected to decrease with each doubling of serum LDL (142 days lower than whites).
	10. Based on a test of the parameters for race and the log2 transformed serum ldl to investigate whether race modifies the relationship between time to death and serum ldl, it appears that the null hypothesis that race is not an effect modifier cannot be rejected (p=0.1034), indicating that race could be acting to modify the relationship observed between time to death and serum ldl but it could also not be acting as a modifier – further studies would be needed to obtain better precision.
	11. To test whether there was an association between time to death and serum ldl, we test whether the log transformed ldl parameters are equal to 0. The obtained p-value (p=0.0184) indicates that there may be an association between time to death and serum ldl since we are able to reject the null that there is no association between time to death and serum ldl.
	12. To test whether there was an association between time to death and race, we test whether the slope parameters obtained for each race are equal to 0. The obtained p-value (p=0.3859) indicates that we are unable to reject the null that there is no association between time to death and race and so we cannot conclude anything from this data set about the association between time to death and race.
	13. To test whether there was a difference in time to death distribution between whites and blacks, we test whether the slope parameters obtained for whites and blacks are simultaneously equal to 0. The obtained p-value (p=0.2285) indicates that we are unable to reject the null that there is no time to death distribution difference between whites and blacks. So we cannot conclude anything from this data set about the differences in time to death distribution between whites and blacks, with this data set.
	14. Based on a regression of the monthly salaries of the 123 individuals who started in 1990 and later and who only had up to 1 year between attaining their highest degree and starting at the university, there appears to be evidence of sex discrimination at the university. On average, the mean monthly salary for women is $571 lower than their male counterparts. This estimate is significant at a level of 0.05 (P<0.0001) and so we can reject the null hypothesis that there is not association between sex and mean monthly salary.
	15. Based on a regression of the monthly salaries of the 123 individuals who started in 1990 and later and who only had up to 1 year between attaining their highest degree and starting at the university, there appears to be evidence of sex discrimination at the university. Using the geometric mean, we find that women earned on average 13.351% lower than the men. The estimate is significant at a level of 0.05 (P<0.0001) and so we can reject the null hypothesis that there is no association between sex and the geometric mean of monthly salary.
	16. If you were interested in absolute differences in the mean monthly salary, the regression in part a would be better for providing a direct measure of the differences in salary among men and women. This is easy to interpret; however, the precision is lower. Since salaries are often calculated on a percent scale, this estimate would break down at much higher or much lower salaries. If, however, the scientific question lies in the relative difference in salaries between men and women, modeling the geometric mean would be better and would provide better precision for the analysis.
	17. If the analyses had not been done correctly, the estimates would still have been the same; however the models would have been less conservative about the estimates as is evident from the wider confidence intervals. When considering associations by year, the estimates would be readjusted such that the models would have become more conservative (narrower confidence intervals).