**Question 1a**

*Methods*: A regression model using a step function was fit as an alternative model to test for a linear relationship between LDL and time to death. The log hazard function was used, where time to death was the dependent variable and LDL was the independent. This step function used breaks at LDL = 0, 70, 100, 130, 160, and 400. The Wald test statistic was used with a two-sided p value to test for departures from linearity.

*Inference*: The coefficient for LDL as a step was -0.001726 (HR 0.9983), with a 95% confidence interval for the hazard ratio as 0.9962 to 1.00. However based on the Wald test statistic, where p = 0.1011, we fail to reject the null that there was no departure from linearity.

**Question 1b**

*Methods*: A regression model using a quadratic polynomial was fit as an alternative model to test for a linear relationship between LDL and time to death. The log hazard function was used, where time to death was the dependent variable and LDL was the independent. The variables LDL and (LDL)2 were used to create the polynomial. The Wald test statistic was used with a two-sided p value to test for departures from linearity.

*Inference*: The coefficient for LDL was -0.02611 (HR 0.9742), with a 95% confidence interval for the hazard ratio as 0.9529 to 0.996. The coefficient for LDL squared was -0.00007631 (HR 1.0001), with a 95% confidence interval for the hazard ratio as 1.00 to 1.00. Based on the Wald test statistic, where p = 0.00248, we reject the null that there was no departure from linearity.

**Question 1c**

*Methods*: A regression model using a cubic polynomial was fit as an alternative model to test for a linear relationship between LDL and time to death. The log hazard function was used, where time to death was the dependent variable and LDL was the independent. The variables LDL, (LDL)2 and (LDL)3 were used to create the polynomial. The Wald test statistic was used with a two-sided p value to test for departures from linearity.

*Inference*: The coefficient for LDL was -0.04221 (HR 0.9587), with a 95% confidence interval for the hazard ratio as 0.9086 to 1.012. The coefficient for LDL squared was 0.0002242 (HR 1.002), with a 95% confidence interval for the hazard ratio as 0.9998 to 1.001. The coefficient for LDL cubed was -0.0000004087 (HR 1.00), wit ha 95% confidence interval for the hazard ratio as 1.00 to 1.00. Based on the Wald test statistic, where p = 0.003328, we reject the null that there was no departure from linearity.

**Question 1d**

*Methods*: A regression model using linear splines was fit as an alternative model to test for a linear relationship between LDL and time to death. The log hazard function was used, where time to death was the dependent variable and LDL was the independent. The linear splines used knots at LDL = 70, 100, 130 and 160. The Wald test statistic was used with a two-sided p value to test for departures from linearity.

*Inference*: The coefficient for the first LDL spline before 70 was -1.2998 (HR 0.2726), with 95% confidence interval 0.04566 to 1.6274. The coefficient for the second LDL spline between 70 and 100 was -1.9273 (HR 0.1455), with 95% confidence interval 0.02979 to 0.7110. The coefficient for the third LDL spline between 100 and 130 was -1.9546 (HR 0.1416), with 95% confidence interval 0.02801 to 0.7161. The coefficient for the fourth LDL spline between 130 and 160 was -2.0130 (HR 0.1336), with 95% confidence interval 0.02597 to 0.6872. The coefficient for the fifth LDL spline after 160 was -2.5459 (HR 0.0784), with 95% confidence interval 0.006589 to 0.9330. Based on the Wald test statistic, where p = 0.01293, we reject the null that there was no departure from linearity.

**Question 1e**

*Methods*: A regression model using a logarithmic transformation was fit as an alternative model to test for a linear relationship between LDL and time to death. The log hazard function was used, where time to death was the dependent variable and log transformed LDL was the independent. The Wald test statistic was used with a two-sided p value to test for departures from linearity.

*Inference*: The coefficient for log transformed LDL was -0.8266 (HR 0.4375), with a 95% confidence interval for the hazard ratio as 0.2738 to 0.6992. Based on the Wald test statistic, where p = 0.000548, we reject the null that there was no departure from linearity.

**Question 1f**

*Methods*: The fitted values from each of the regression fits in Question 1a – 1e were plotted, as well as the model with only centered serum LDL values. These fitted values were used to compare the similarities and differences among the models.

*Inference*: The five models from the previous segments, and the model with centered LDL values are very similar. This is shown below in Figure 1. The fitted survival probabilities for each model are close across all serum LDL values. However, the model fit from part A, for a step function, had slightly lower survival probabilities. It is notable that this was the only model where the test statistic suggested there was no evidence to reject the null of no departure from linearity. There were no visible differences between the other models’ survival probabilities.



*Figure 1. Plot of the Fitted Values from Each Model*

**Question 2a**

A model with linear splines allows for a piecewise polynomial representation of serum LDL as it is associated with time to death. By allowing for these piecewise components, the model more accurately fits the association of LDL to time of death. Without the linear splines, the fit of LDL would not be sensitive to the piecewise changes between intervals. However, the splines also allow information to be borrowed across all intervals, making it more powerful than simply using dummy variables. The output of the model provides regression parameters for each spline segments. Interpreting these parameters is similar to dummy variables, as each coefficient takes into account the other piecewise components.

**Question 2b**

There is no evidence that the association between time to death and serum LDL is truly U-shaped. This is best illustrated by the plot of time to death by serum LDL, which is shown below in Figure 2. Although there is somewhat of a u-shaped trend, the majority of the values hover across the top in a horizontal line.



*Figure 2. Relationship between Time to Death and LDL*

**Question 3a**

A model was fit of time to death regressed on log transformed LDL, race, and their interaction. The coefficient for log transformed was -0.7744 (HR 0.461), indicating that for each multiplicative increase in LDL survival time decreases. The coefficient for race 2 (black) compared to race 1 (white) was -1.867 (HR 0.1545), indicating that those with black race have lower survival times than white race. The coefficient for race 3 (Asian) compared to race 1 (white) was 5.720 (HR 304.8), indicating that those with Asian race have much higher survival times than white race. The coefficient for race 4 (other) compared to race 1 (white) was 19.62 (HR 3.3x108), indicating that those with other race have much, much higher survival times than white race. The interaction of log transformed LDL for race 2 (black) was 0.4398 (HR 1.552), indicating that each multiplicative increase in LDL for black race increased survival times, as compared to comparable multiplicative LDL increases for white race. The interaction of log transformed LDL for race 3 (Asian) was -1.170 (HR 0.3102), indicating that each multiplicative increase in LDL for Asian race decreased survival times, as compared to comparable multiplicative LDL increases for white race. The interaction of log transformed LDL for race 4 (other) was -4.021 (HR 0.01793), indicating that each multiplicative increase in LDL for other race decreased survival times, as compared to comparable multiplicative LDL increases for white race.

**Question 3b**

The regression analysis from part A suggests that race does not modify the association between time to death and serum LDL. This hypothesis can be tested by looking at the p-values for the interaction terms, where none of them are significant at alpha = 0.05. Therefore, since no interaction term is significant, we can conclude that race does not modify the association between time to death and serum LDL, for p < 0.05. Furthermore, an F test between the model from part A and a model without race included suggests no statistical difference between the models (p = 0.3394).

**Question 3c**

The regression analysis from part A can be used to test the hypothesis that there is no association between time to death and serum LDL. The parameter for log transformed LDL was significant at p = 0.00447. Therefore we can reject this hypothesis and conclude there is no evidence to suggest that there is no association between time to death and serum LDL.

**Question 3d**

The regression analysis from part A can be used to test the hypothesis that there is no association between time to death an race. There were three regression parameters included for race (black referenced to white; Asian referenced to white; other referenced to white). The parameter for black was not significant (p=0.636), the parameter for Asian was not significant (p=0.21769), however the parameter for other was significant (p=0.04890). Since the other race was significant, we can reject the hypothesis and conclude there is no evidence to suggest that there is no association between time to death and race.

**Question 3e**

The regression analysis from part A can be used to test the hypothesis that there is no association between the distribution of time to death between whites and blacks. The parameter for race black was in comparison to white race. This variable was not significant in the model, p = 0.636. Therefore, we fail to reject the hypothesis that there is no association between the distributions of time to death between whites and blacks.

**Question 4a**

*Methods*: A linear regression model was fit on salary by sex, year and the interaction of sex and year, to determine if there was evidence of sex discrimination in the mean salary in recent years.

*Inference*: There is no evidence of sex discrimination in the mean salaries for the recent years. This is supported by the variable of sex and interaction term, neither of which are significant at alpha = 0.05. Gender had a p-value of 0.9248 and gender by year had a p-value of 0.99059. Therefore, we fail to reject the null hypothesis that gender does not affect mean salary.

**Question 4b**

*Methods*: Similar to before, a linear regression model was fit on salary by sex, year, and the interaction of sex and year. However, this time salary was log transformed to determine if there was evidence of sex discrimination in the geometric mean salary in recent years.

*Inference*: There is no evidence of sex discrimination in the geometric mean salaries for the recent years. This is supported by the variable of sex and interaction term, neither of which are significant at alpha = 0.05. Gender had a p-value of 0.677 and gender by year had a p-value of 0.747. Therefore, we fail to reject the null hypothesis that gender does not affect geometric mean salary.

**Question 4c**

The two models are adequate for use in answering the question of sex discrimination in mean salary. However the model from part B, on the geometric mean of salary, has more merit. This is because analyzing the multiplicative increases in salary is more meaningful than the unit increases of a linear model (a difference of a few dollars would not be of much interest).

**Question 4d**

The dataset includes multiple observations for several employees; therefore there is correlation between several observations. If the analysis in parts A and B were done incorrectly by treating the data as independent, the models would have yielded anticonservative inferences about associations with sex. However, there would not be any difference between the models if considering associations by year, because then there would only be one observation per subject per year.