**Question 1.**

1. Yes this is a saturated model. This is modeling 4 groups with 3 predictors and the incercept.

**Methods:** Logistic regression was performed to investigate an association between race and diagnosed diabetes prevalence. Race was categorized into four groups: white, Black, Asian, and other. Race was modeled using ANOVA (dummy variables), with white as the reference group. Odds ratios were calculated comparing the prevalence of diabetes across race groups, with confidence intervals and two-sided p-values computed using Wald statistics and robust standard errors (Huper-White sandwhich estimator).

**Inference:** All 735 had information on race and diabetes status. 78% of subjects were white, 14% were Black, 6% were Asian, and 2% were another race. 79 (10.8%) subjects had diabetes at baseline. From the logistic regression analysis, based on a two-sided p-value of 0.0956, we found that overall there was insufficient evidence to reject the null hypothesis that diabetes odds is the same across racial groups. There was insufficient evidence to support an association between the odds of diabetes and race over all.

1. **Intercept:** The log odds of having diabetes among whites was -2.22 (95% CI: -2.50, -1.94). This corresponds to an odds of 0.109. Based on a 95% confidence interval, this odds of diabetes among whites would not be unusual if the true odds of diabetes among whites were between 0.0824 and 0.143.

**Race2:** The odds of having diabetes among Blacks was 1.93 times the odds of having diabetes among whites. Based on a 95% confidence interval, this odds ratio of 1.93 would not be unusual if the true odds ratio comparing the odds of diabetes among blacks to that of whites were between 1.08 and 3.44.

**Race3:** The odds of having diabetes among Asians was 0.628 times the odds of having diabetes among whites. Based on a 95% confidence interval, this odds ratio of 0.628 would not be unusual if the true odds ratio comparing the odds of diabetes among Asians to that of whites were between 0.189 and 2.09.

**Race4:** The odds of having diabetes among other races was 1.84 times the odds of having diabetes among whites. Based on a 95% confidence interval, this odds ratio of 1.84 would not be unusual if the true odds ratio comparing the odds of diabetes among other races to that of whites were between 0.393 and 8.63.

1. If we were to ignore the issue of multiple comparisons, using a 0.05 level of significance, we would conclude that there is sufficient evidence to reject the null hypothesis that Blacks and whites have the same odds of diabetes, based on a two-sided p-value of 0.026. Thus, we can reject the null hypothesis that Blacks have the same odds of diabetes as whites, in favor of blacks having greater odds of diabetes compared to whites. There remains insufficient evidence to reject the null hypothesis that Asians and other races have the same odds of diabetes as whites, based on their two-sided p-values of 0.449 and 0.438 respectively.
2. Performing the same logistic regression as above, only using Black race as the reference group, would result in an identical inference as reported in part a, with the same two-sided p-value of 0.0956. The same conclusion of failure to reject the null hypothesis and detect an association between race and diabetes applies here. This is a reparameterization of the saturated model used in part a, so the overall inference remains the same, regardless of the reference group used. What changes are the individual interpretations of the parameters, because the reference group is different.
3. **Intercept:** The log odds of having diabetes among Blacks was -1.56 (95% CI: -2.07, -1.06). This corresponds to an odds of 0.209. Based on a 95% confidence interval, this odds of diabetes among blacks would not be unusual if the true odds of diabetes among blacks were between 0.126 and 0.348.

**Race1:** The odds of having diabetes among whites was 0.519 times the odds of having diabetes among Blacks. Based on a 95% confidence interval, this odds ratio of 0.519 would not be unusual if the true odds ratio comparing the odds of diabetes among whites to that of blacks were between 0.291 and 0.925.

**Race3:** The odds of having diabetes among Asians was 0.326 times the odds of having diabetes among Blacks. Based on a 95% confidence interval, this odds ratio of 0.326 would not be unusual if the true odds ratio comparing the odds of diabetes among Asians to that of Blacks were between 0.0909 and 1.17.

**Race4:** The odds of having diabetes among other races was 0.956 times the odds of having diabetes among Blacks. Based on a 95% confidence interval, this odds ratio of 0.956 would not be unusual if the true odds ratio comparing the odds of diabetes among other races to that of Blacks were between 0.193 and 4.74.

1. If we were to ignore the issue of multiple comparisons, using a 0.05 level of significance, we would conclude that there is sufficient evidence to reject the null hypothesis that whites and Blacks have the same odds of diabetes, based on a two-sided p-value of 0.026. Thus, we can reject the null hypothesis that Blacks have the same odds of diabetes as whites, in favor of blacks having greater odds of diabetes compared to whites. There remains insufficient evidence to reject the null hypothesis that Asians and other races have the same odds of diabetes as blacks, based on their two-sided p-values of 0.085 and 0.956 respectively.
2. The p-values for the comparison of the odds of diabetes between whites and blacks remains the same (0.0296). However, the p-values comparing whites to Asians and others and much larger than the p-values comparing blacks to Asians and others. The magnitude of the point estimates also differs depending on the reference group. This shows that using different reference groups can potentially result in different interpretations of the significance of the relationship for different parameters, while the overall p-value remains the same in this saturated model. An individual parameter may appear more significant and meaningful in one model than another depending on the reference group, so using the individual parameter estimates is a somewhat arbitrary measure for making decisions to include the parameter or not in a model.

**Question 2**

1. **Descriptive Statistics:**

**Descriptive Stats Methods:** Descriptive statistics for survival were calculated using Kaplain Meier and include the mean, median and range of follow-up time. Descriptive statistics for serum LDL levels included the number of cases with missing data, and the minimum, maximum, mean, median, and standard deviation. Serum LDL was categorized based on the Mayo Clinic’s guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. Kaplan-Meier estimates of survival were calculated and graphed for each LDL category.

**Descriptive Stats Inference:**

735 subjects were followed for death from any cause. Using Kaplan-Meier, subjects were had an estimated average follow-up time of 5.33 years (median 5.66 years, range 5.00 to 5.91 years). 133 deaths were observed during follow-up. Baseline serum LDL measurements were missing for 10 subjects. For the 725 subjects with serum LDL measurements, the mean LDL was 126 mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL). Table 1 presents the survival distribution estimates by serum LDL category for the 725 subjects with available LDL data (10 subjects had no available baseline LDL information). The 2 year survival probability was similar across LDL groups (all >93%). The 5 year survival probability was similar for all LDL groups over 70 mg/dL (81%-88%), but was much lower for the lowest LDL group (59%). Figure 1 presents the Kaplan-Meier survival probability estimates graphically, where it is again the lowest LDL group that shows the most markedly different survival distribution.

**Table 1. Kaplan-Meier estimates of distribution of time from study enrollment to death from any cause for subjects having baseline serum LDL measurements.**

|  |  |  |
| --- | --- | --- |
|  | **Baseline Serum LDL (mg/dL)** |  |
|  | **11 – 69** | **70 – 99** | **100 – 129** | **130 – 159** | **160 – 189** | **190 – 247** | **All Subjects\*** |
| **N** | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| **Deaths** | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| **2 year Survival Probability** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 96.7% |
| **5 year Survival Probability** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 86.0% |
| **10th %tile of Survival** | 3.46 | 3.8 | 3.41 | 4.3 | 4.53 | 4.13 | 3.66 |
| **20th %tile of Survival** | 3.55 | 5.44 | 5.36 | - | - | - | 5.54 |
| **5.75 Year Restricted Mean of Survival** | 4.91 | 5.24 | 5.23 | 5.35 | 5.45 | 5.32 | 5.29 |

**\*10 subjects had no available LDL measurements at baseline and were therefore excluded from all analyses.**

**Figure 1. Kaplan-Meier based survival estimates of distribution of time from study enrollment to death from any cause (n=725).**



**Inferential Statistics**

**Methods:** Cox proportional hazards regression was performed comparing the distribution of time to death across baseline serum LDL level categories. Serum LDL was categorized based on the Mayo Clinic’s guidelines: less than 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and greater than or equal to 190 mg/dL. These LDL categories were modeled using ANOVA, with LDL categories defined as dummy variables. The lowest LDL category (less than 70 mg/dL) was used as the reference group. The association between LDL and mortality was summarized by regression estimated hazard ratios. Wald statistics with the Huber-White sandwich estimator (robust standard errors) were used to calculate 95% confidence intervals and two-sided p-values. Subjects missing baseline LDL data were excluded.

**Inference:** 725 subjects had baseline LDL data and were included in the analysis. 131 of the included subjects were observed to die during the follow-up period with an average observation time of 5.33 years. From cox proportional hazards regression analysis, we found an overall two-sided p-value of 0.0087. Based on a significance level of 0.05, we can reject the null hypothesis that the instantaneous risk of death is the same across all defined categories of LDL.

1. **Intercept:** There is no intercept in cox proportional hazard regression. This is a semi-parametric model in which the hazard ratio is the parameter(s) of interest. The baseline hazard function is unspecified, because it is a function over time and is therefore not constant.

**IldlCTG\_70**: The instantaneous risk of death (hazard) among those with serum LDL bewteen 70 and 99 mg/dL was a relative 60.2% lower (hazard ratio: 0.398; 95% CI: 0.203, 0.782) than the hazard of those with a serum LDL less than 70 mg/dL. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true hazard of those with LDL between 70 and 90 mg/dL were between 21.8% and 79.7% lower than that of those with LDL less than 70 mg/dL.

**IldlCTG\_100:** The instantaneous risk of death (hazard) among those with serum LDL bewteen 100 and 129 mg/dL was a relative 60.7% lower (hazard ratio: 0.393; 95% CI: 0.207, 0.744) than the hazard of those with a serum LDL less than 70 mg/dL. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true hazard of those with LDL between 70 and 90 mg/dL were between 25.6% and 79.3% lower than that of those with LDL less than 70 mg/dL.

**IldlCTG\_130:** The instantaneous risk of death (hazard) among those with serum LDL bewteen 130 and 159 mg/dL was a relative 70.6% lower (hazard ratio: 0.294; 95% CI: 0.152, 0.568) than the hazard of those with a serum LDL less than 70. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true hazard of those with LDL between 70 and 90mg/dL were between 43.2% and 84.8% lower than that of those with LDL less than 70 mg/dL.

**IldlCTG\_160:** The instantaneous risk of death (hazard) among those with serum LDL bewteen 160 and 189 mg/dL was a relative 74.3% lower (hazard ratio: 0.257; 95% CI: 0.113, 0.580) than the hazard of those with a serum LDL less than 70. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true hazard of those with LDL between 70 and 90mg/dL were between 42.0% and 88.7% lower than that of those with LDL less than 70 mg/dL.

**IldlCTG\_190:** The instantaneous risk of death (hazard) among those with serum LDL bewteen 160 and 189 mg/dL was a relative 68.3% lower (hazard ratio: 0.317; 95% CI: 0.101, 0.989) than the hazard of those with a serum LDL less than 70. Based on a 95% confidence interval, this observed hazard ratio would not be unusual if the true hazard of those with LDL between 70 and 90mg/dL were between 1.10% and 89.9% lower than that of those with LDL less than 70 mg/dL.

1. The likelihood ratio test can be used to assess which model provides a better fit. The results of the analysis based on a two-sided p-value of 0.4776, was that we reject the null that the models differ in their ability fit the data. However, the overall Wald statistic using robust standard errors was 6.76 when modeling LDL as a continuous variable, and was 15.42 when modeling LDL as dummy variables, indicating an association was better detected when using the dummy variable model.
2. See question 4

**Question 3**

**a)** Refer to descriptive statistics methods and inference for question 2 above**.**

**Inferential Statistics**

**Methods:** Cox proportional hazards regression was performed comparing the distribution of time to death across baseline serum LDL levels defined using linear splines. Serum LDL had knots based on the Mayo Clinic’s guidelines: at 70, 100, 130, 160, and 190 mg/dL. The association between LDL and mortality was summarized by regression estimated hazard ratios. Wald statistics with the Huber-White sandwich estimator (robust standard errors) were used to calculate 95% confidence intervals and two-sided p-values. Subjects missing baseline LDL data were excluded.

**Inference:** 725 subjects had baseline LDL data and were included in the analysis. 131 of the included subjects were observed to die during the follow-up period with an average observation time of 5.33 years. From cox proportional hazards regression analysis modeling LDL as linear splines, we found an overall two-sided p-value <0.0001. Based on a significance level of 0.05, we can reject the null hypothesis that the instantaneous risk of death is the same across all levels of LDL.

**b)** **Intercept:** There is no intercept in cox proportional hazard regression. This is a semi-parametric model in which the hazard ratio is the parameter(s) of interest. The baseline hazard function is unspecified, because it is a function over time and is therefore not constant.

**ldl0:** The estimated instantaneous risk of death is a relative 19.9% lower (HR: 0.801, 95% CI: 0.666, 0.964) for each 10 mg/dL higher serum LDL level at baseline among those with a baseline LDL less than 70 mg/dL. This hazard ratio would not be unusual if the true hazard for each 10 mg/dL increase in LDL were between 3.61% and 33.4% lower for those with LDL less than 70 mg/dL.

**Ld70:** The estimated instantaneous risk of death is a relative 18.5% lower (HR: 0.815, 95% CI: 0.621, 1.069) for each 10 mg/dL higher serum LDL level at baseline among those with a baseline LDL between 70 and 99 mg/dL. This hazard ratio would not be unusual if the true hazard for each 10 mg/dL increase in LDL were between 37.9% lower and 6.91% higher for those with LDL between 70 and 99 mg/dL.

**Ld100:** The estimated instantaneous risk of death is a relative 2.27% lower (HR: 0.977, 95% CI: 0.788, 1.213) for each 10 mg/dL higher serum LDL level at baseline among those with a baseline LDL between 100 and 129 mg/dL. This hazard ratio would not be unusual if the true hazard for each 10 mg/dL increase in LDL were between 21.2% lower and 21.3% higher for those with LDL between 100 and 129 mg/dL.

**ld130:** The estimated instantaneous risk of death is a relative 3.67% higher (HR: 1.037, 95% CI: 0.812, 1.324) for each 10 mg/dL higher serum LDL level at baseline among those with a baseline LDL between 130 and 159 mg/dL. This hazard ratio would not be unusual if the true hazard for each 10 mg/dL increase in LDL were between 18.8% lower and 32.4% higher for those with LDL between 130 and 159 mg/dL.

**ld160:** The estimated instantaneous risk of death is a relative 25.6% lower (HR: 0.744, 95% CI: 0.483, 1.147) for each 10 mg/dL higher serum LDL level at baseline among those with a baseline LDL between 160 and 189 mg/dL. This hazard ratio would not be unusual if the true hazard for each 10 mg/dL increase in LDL were between 51.7% lower and 14.7% higher for those with LDL between 160 and 189 mg/dL.

**ld190:** The estimated instantaneous risk of death is a relative 32.8% higher (HR: 1.328, 95% CI: 0.810, 2.179) for each 10 mg/dL higher serum LDL level at baseline among those with a baseline LDL greater 190 mg/dL or greater. This hazard ratio would not be unusual if the true hazard for each 10 mg/dL increase in LDL were between 19.0% lower and 2.179 times higher for those with LDL 190 mg/dL or greater.

**c)** The likelihood ratio test can be used to assess which model provides a better fit. The results of the analysis based on a two-sided p-value of 0.3730, was that we reject the null that the models differ in their ability fit the data. However, the overall Wald statistic using robust standard errors was 6.76 when modeling LDL as a continuous variable, and was 31.77 when modeling LDL as linear splines, indicating an association was better detected when using the linear spline model.

**d)** See question 4

**Question 4:**

**a)** The strategies used in homeworks 4 and 5 have the advantage of using the data more completely. In homeworks 1-3, the outcome was defined as a binary indicator of 5 year survival, whereas in homeworks 4 and 5, the instantaneous risk of death (hazard) was used, which allows for gaining additional information from the data, and compares survival over time, not only at one the one time point at 5 years. Additionally, defining LDL as a binary predictor in homeworks 1-3 is potentially problematic as the groups may have additional variation within the 2 categoties that is lost in this type of model. Defining LDL as a continuous linear variable, or a logorythmic variable, only allows for 1 distinct shape, where the direction may not always be consistent. Dummy variables have the advantage of comparing multiple defined groups which fit exactly with the parameter estimates when no other predictors are in the models. They do not have to have a linear relationship because they ignore the order of the predictor that has been categorized, which allows for greater flexibility when investigating the relationship, which may not be linear, or even in the same direction. The linear spline model allows for the greatest flexibility and retains the most information, in that LDL remains continuous using all of the LDL data, and also allows for changes in the slope at specific knots if the relationship does not fit a linear or logarithmic model. It also allows for changes in the direction of the relationship, which is more flexible than other models that defined LDL continuously.

**b)** In all models, there is an overall downward trend in the hazard with increasing LDL. The dummy variable model is similar to the 3 models from homework 4, matching the u-shape model well at the higher LDL levels. However, the extremes of the LDL levels have fewer data points, so it is difficult make conclusions regarding model fit. The linear spline model has a much lower hazard ratio than the other models for all levels of LDL.

****

**c)** I wouldprefer a using the dummy variables model for the instantaneous risk of death. Using multiple categories is superior to dichotomizing LDL in terms of precision gained, though it is worse than using LDL continuously in terms of precision. However, because LDL is defined into the 6 categories by the Mayo clinic, the risk of death between these groups is of scientific interest, and the compared risk of death within these categories is meaningful in practice. Using the instantaneous risk of death as opposed to 5 year survival probability is better because it uses more of the data, and will estimate if certain categories of LDL increase the risk of death overall.