In problems 2 through 4, we restrict our analysis to the elderly observations in the dataset with non-missing values of serum LDL; this leaves 725 of the original 735 observations.

**(1)**

**(a)**

**Methods:** We fit the following logistic regression model

,

where is an indicator of whether the *i*th observation was diagnosed with diabetes before the MRI, is an indicator of whether the *i*th individual is black, is an indicator of whether the *i*th individual is Asian, and is an indicator of whether the *i*th individual belong to another race. We use a Hubert-White sandwich estimator of standard errors to acknowledge the possibility of unequal variances across groups. We also compute a Wald-based confidence interval for these estimated odds ratios, associated two-sided p-values (for the coefficient estimates), and a Wald chi-squared test statistic to assess the joint association between the entire set of race categories and the odds of a diabetes diagnosis.

**Inference:**

This is a saturated model, since we have four groups (white, black, Asian, and other) along with 4 parameters (three slope parameters and one intercept).

Since interpretation of the 3 slope and 1 intercept parameters in our model is complicated by the multiple testing problem, we instead use the Wald chi-squared test (with 3 degrees of freedom) to measure an association between the entire set of race categories and the odds of a diagnosis of diabetes in our data. Here, we have a Wald chi-squared test statistic of 6.35 with an associated p-value of 0.096 > 0.05. We do not have evidence to reject the null hypothesis that the odds of a diabetes diagnosis is not associated with race in our data.

**(b)**

The exponentiated slope estimates report the odds ratios for diabetes in race categories (relative to the omitted category of white individuals) as defined in part (a). In addition, will estimate the log odds of diabetes in white individuals, so ) reports the estimated odds of a diabetes diagnosis for white individuals, ) estimates the diabetes diagnosis odds ratio for black individuals relative to whites, ) estimates the diabetes diagnosis odds ratio for Asian individuals relative to whites, and ) estimates the diabetes odds ratio for observations belonging to the “other” racial group relative to whites.

**(c)**

If we ignored the issues of multiple comparisons, we would conclude that the estimated odds of a diabetes diagnosis among white individuals is significantly different from 1, as ) has a point estimate of 0.109 with a 95% Wald CI of [0.082, 0.143] and two-sided p-value . We would also conclude that the odds of a diabetes diagnosis among black individuals is significantly higher than it is among white individuals, as ) has an odds ratio point estimate of 1.929 with a 95% Wald CI of [1.054, 3.447] and two-sided p-value of 0.033 < 0.05.

**(d)**

If we fit a logistic regression model with blacks as the reference group, we would just be re-parameterizing the regression model we describe in part (a). In the guiding equation in part (a), we would now have as an indicator of whether the *i*th individual is white. Due to the same multiple testing problem, we would still be interested in the Wald chi-squared test (with 3 degrees of freedom) to measure an association between the entire set of race categories and the odds of a diagnosis of diabetes in our data. The result of the Wald chi-squared test would be the same as it was with whites as the reference group since the joint set of race categories in unchanged.

**(e)**

Due to the re-parameterization of our model, we would obtain different slope and intercept estimates than in part (b). Now ) will report the estimated odds of a diabetes diagnosis for black individuals, ) estimates the diabetes diagnosis odds ratio for white individuals relative to blacks, ) estimates the diabetes diagnosis odds ratio for Asian individuals relative to blacks, and ) estimates the diabetes odds ratio for observations belonging to the “other” racial group relative to blacks.

**(f)**

Had we ignored the multiple testing problem, we would conclude that the estimated odds of a diabetes diagnosis among black individuals is significantly different from 1, as ) has a point estimate of 0.209 with a 95% Wald CI of [0.126, 0.348] and two-sided p-value . We would also conclude that the odds of a diabetes diagnosis among white individuals is significantly lower than it is among white individuals, as ) has an odds ratio point estimate of 0.519 with a 95% Wald CI of [0.291, 0.925] and two-sided p-value of 0.026 < 0.05. We also see that the estimate of the diabetes odds ratio for Asians relative to blacks has a p-value that falls just shy of statistical significant at the 0.05 level (0.085) with an associated point estimate of 0.326 and 95% Wald CI of [0.091, 1.167]. In our original specification, the coefficient estimates for the Asian and other race categories both had p-values larger than 0.40. Although we reach the same conclusions in this inference as in part (d) at the 0.05 level of significance, the relative magnitudes of the p-values for the Asian and other racial groups have changed substantially.

**(g)**

These results suggest that “stepwise model building” runs the risk of excluding a dummy variable that can improve the overall fit of the model, if its inclusion is determined by only its individual coefficient estimate. We have seen that the individual p-values of the coefficient estimates on each dummy variable can vary considerably based only on re-parameterizations of the model while contributing the same level of aggregate information to our association testing. A more reliable criterion for assessing the importance of dummy variables would their collective explanatory power in the Wald chi-squared test of their joint significance.

In addition, we found that the diabetes prevalence does not vary significantly across the joint set of race categories, but it was found to be significant when we only tested whether it varied between white observations and black individuals. It may not be wise to include a dummy variable indicating white/black race if we have found that the entire set of racial indicators does not significantly improve the model’s fit.

**(2)**

**Methods:** We present descriptive statistics and survival distributions using Kaplan-Meier estimates across 6 strata based on serum LDL level based on the recommendations from the Mayo Clinic used for Homework 1 (< 70 mg/dL, 70-99 mg/dL, 100-129 mg/dL, 130-159 mg/dL, 160-189 mg/dL, and ≥190 mg/dL).We estimate hazard ratios using Cox proportional hazard regression with a Hubert-White sandwich estimator of corresponding standard errors and a Wald-based chi-squared test statistic to evaluate the significance of chosen covariates. Our explanatory variables are dummy variables of an observation’s Mayo Clinic LDL risk category. Our response variable is an indicator of all-cause mortality.

**Descriptive Statistics:** The lowest LDL group (< 70 mg/dL) had the lowest estimate of 5-year survival, at 59.1%. The estimated 5-year survival probably was between 80% and 88% for the 5 other LDL risk categories. The lowest LDL group also had the shortest estimated length of life during the study period, at 4.91 years, compared to an estimate of above 5 years for the other groups. The Kaplan-Meier survival curves also show that the estimated survival distribution of the lowest LDL group begins to differ substantially from the other groups beginning at around 1250 days and remains noticeably lower for the duration of the study period. The lowest LDL group has the small sample size of any group (22 observations), although the largest LDL group is close with only 24 observations. The maximum observed time in the study is 5.91 years and the minimum is 68 days.



Figure 1: Kaplan-Meier Survival Curves by LDL group

|  |
| --- |
| **Estimated Probability of Survival by Mayo Clinic LDL Risk Category (N=725)** |
|  | **<70 mg/dL** **(N=22)** | **70-100 mg/dL (N=143)** | **100-130 mg/dL (N=228)** | **130-160 mg/dL****(N=225)** | **160-190 mg/dL****(N=83)** | **≥190 mg/dL****(N=24)** |
| **Subjects** | 22 | 143 | 228 | 225 | 83 | 24 |
| **Deaths** | 10 | 28 | 44 | 34 | 11 | 4 |
| **2-Year Survival Rate** | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% |
| **5-Year Survival Rate** | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% |
| **Restricted Mean****(Years)** | 4.91 | 5.24 | 5.23 | 5.35 | 5.45 | 5.32 |

**Inference:** From Cox proportional hazards regression with a Hubert-White estimator of standard errors, we find evidence that instantaneous mortality risk is associated with our 6-level LDL risk category at the 0.05 level of significance. We estimate a Wald chi-squared test statistic (with 5 degrees of freedom) of 15.42, with an associated one-sided p-value of 0.009 < 0.05. As a result, we conclude that our observed data are not consistent with identical hazard between groups, or hazard ratios of 1 across 5 levels of our risk categories relative to the omitted category.

**(b)**

The omitted LDL risk category in our regression model is the lowest LDL group (<70 mg/dL). Each of the remaining 5 risk categories has an estimated hazard ratio that reports the instantaneous mortality in that category relative to the lowest LDL category. We estimate that observations with between 70-99 mg/dL have a mortality risk that is 60.2% lower than the group with LDL below 70 mg/dL; individuals with 100-129 mg/dL have an estimated mortality risk that is 60.7% lower than the lowest LDL group; those with 130-159 mg/dL have a mortality risk of 70.6% lower; those with 160-189 mg/dL have a mortality risk of 74.4% lower; and those with LDL of at least 190 mg/dL are estimated to have a mortality risk that is 68.% lower than those in the lowest LDL group. The intercept reports the hazard of the lowest LDL group relative to itself, so it will be 100%.

**(c)**

We can fit a regression model that includes both dummy variables for the 6-level LDL risk categories as well as a continuous measure of LDL (with the lowest category omitted as a reference group) to test the performance of the model in this problem. We then test the null hypothesis that instantaneous mortality risk is not associated with the categorical measure of LDL risk categories in the presence of a continuous measure of LDL. We perform this on our data using a Wald test of the linear null hypothesis and find a chi-squared test statistic (based on 5 degrees of freedom) of 5.14 with an associated p-value of 0.399. As a result, we do not have sufficient evidence to reject the null hypothesis of each level of the dummy variable being unassociated with instantaneous mortality risk. We thus have some evidence that the model in this problem does not provide a better fit than a regression model whose only explanatory variable is a continuous measure of LDL.

**3.**

**(a)**

**Methods:** We estimate the association between LDL risk level (as defined in Question 2) and all-cause mortality in our sample. We estimate hazard ratios using Cox proportional hazard regression with a Hubert-White sandwich estimator of corresponding standard errors. We fit a piecewise linear function of LDL with 5 knots separating observations into the 6 Mayo Clinic levels of LDL risk level. The slopes of these six within-group linear components comprise our explanatory variables. These allow us to fit a potential non-linear association between LDL and instantaneous mortality (if the slopes are not equal across groups) using a Wald test. Our response variable is an indicator of all-cause mortality.

**Descriptive Statistics:** See Question 2(a) for Kaplan-Meier descriptive statistics on all-cause mortality for individuals in the six LDL risk categories.

**Inference:** Our main inference will be a test for linearity in our proportional hazards model (i.e., that the slopes of each of the 6 components of the piecewise linear fit are equivalent) using a post-estimation Wald test. I describe this in part (c). In a test for association, interpretation of the 6 slope parameters in our model is complicated by the multiple testing problem, so we are instead interested in the Wald chi-squared test (with 6 degrees of freedom) to measure an association between the entire set of LDL risk categories and instantaneous mortality risk. Here, we have a Wald chi-squared test statistic of 31.77 with an associated p-value of. So we have evidence of an association between all-cause hazard and LDL risk category under this model.

**(b)**

Each of the 6 slope parameters in our model (one for each of the 6 LDL risk categories) corresponds to the estimated difference in instantaneous risk of mortality associated between groups of observations with LDLs that differ by 1 mg/dL *within the given risk category.* For example, the slope parameter for the lowest LDL group estimates the difference in instantaneous mortality risk for two groups of observations with LDLs below 70 mg/dL who have LDLs that differ by 1 mg/dL. The intercept would estimate the instantaneous rate of death for an individual with LDL of 0 mg/dL, so it is not scientifically meaningful for our problem.

**(c)**

We test for linearity in our proportional hazards model (i.e., that the slopes of each of the 6 components of our piecewise linear fit are equivalent) using a post-estimation Wald test. Our null hypothesis is that the piecewise linear slopes across our 6 LDL risk categories are identical. We find a chi-squared test statistic (with 5 degrees of freedom) in our data of 9.88, with an associated p-value of 0.079 > 0.05. So at the 0.05 level of significance, we do not have sufficient evidence to conclude that our observed data are inconsistent with a straight-line fit of instantaneous mortality risk as a function of LDL risk category, using the Wald test of linearity.

**4.**

**(a)**

Our approaches in Homework 4 and 5 have allowed us to test for non-linearity in the association between LDL and mortality. If the true trend between these variables is non-linear, these techniques may allow us to fit a model that provides more reliable inference from our data. The approaches in HW 1-3 did not allow for us to do this. In addition, we are no longer dichotomizing the LDL variable into just a “high” and “low” category as we did in HW 1-2. Instead, we are using finer levels (6 in this homework) that yield a categorization of LDL that may have more scientific importance and explanatory power. In HW 4-5, we do not restrict ourselves to the assumption of equal variance between levels of explanatory variables, as we had sometimes done in previous assignments. We have also begun looking at mortality over the entire length of the study period instead of only over the first 5 years, which now lets us use some techniques designed for dealing with censored data.

**(b)**

Figure 2 plots the fitted values from the three proportional hazards models in HW 4 and the proportional hazards models fit in problems 2 and 3 from this homework. The estimated hazard ratios are relative to a reference group with an LDL of 160 mg/dL. The fitted values appear to be quite similar for values of LDL above roughly 75 mg/dL and less than 220 mg/dL, estimating an overall trend of decreasing hazard ratio for increasing LDL in this range. We see some slight differences between the magnitude of the upward trend fitted values for observations with the largest levels of LDL (roughly above 220 mg/dL), although it is not too dramatic except for the individual with the highest LDL in the dataset (247 mg/dL).We cannot tell if this upward trend is meaningful without having more data points in this region of highest LDLs.

The most noticeable differences appear to be in the fitted values for individuals with LDL below 70 mg/dL, which correspond to the lowest LDL group in both HW4 and HW5. The two analyses from HW 5 appear to estimate the highest fitted values for these observations with the lowest LDL. This may be because we used finer LDL categories in HW 5 (six instead of three) and are able to estimate true relative hazard more precisely as a result, but we cannot be sure of this without a considerably larger sample size and more detailed analysis. The linear splines fit shown in yellow points appears piecewise linear as expected.



**(c)**

*A priori,* I would have preferred the linear model of an association between instantaneous mortality risk and a continuous measure of LDL specified in problem 1 of HW 4. I would prefer not to categorize the continuous LDL variable so that we do not lose any information about its distribution. I would prefer not to rely on recommendations for LDL risk categories that may not have been tested scientifically. I think that pre-specifying the exact knots for the linear splines model according to these recommendations is too strong of an assumption. I think it is reasonable to expect a general linear trend to satisfactorily answer our guiding scientific question, while interpretation of the parameter estimates from some of our non-linear models may not be as straightforward. The non-linear models also run the risk of over-fitting our data and since we do not have a particularly large sample (particularly near the boundaries of observed LDL) I would be most comfortable just using our data to test for a linear association between continuous LDL and instantaneous risk of death.