In all cases, 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 fit for analysis.

**34/40**

**(1)**

**Methods:** We present descriptive statistics and survival distributions using Kaplan-Meier estimates across 3 strata based on serum LDL level (<100 mg/dL, 100-159 mg/dL, and ≥160 mg/dL).. We define these strata based on the recommendations from the Mayo Clinic used for Homework 1 to correspond roughly to low, medium, and high levels of LDL. We compute a hazard ratio and accompanying Wald-based 95% CI using Cox proportional hazard regression with a Hubert-White sandwich estimator of corresponding standard errors. Our explanatory variable is a continuous measure of LDL. Our response variable is an indicator of all-cause mortality.



Figure 1: Kaplan-Meier Survival Curves by LDL group

|  |
| --- |
| **Estimated Probability of Survival by LDL Group** |
| **Year** | **< 100 mg/dL LDL****(N=165)** | **100-159 mg/dL LDL****(N=453)** | **≥ 160 mg/dL LDL****(N=107)** |
| 1 | 0.982 | 0.980 | 1.00 |
| 2 | 0.964 | 0.947 | 0.981 |
| 3 | 0.909 | 0.921 | 0.953 |
| 4 | 0.867 | 0.894 | 0.907 |
| 5 | 0.800 | 0.841 | 0.869 |

**Inference:** The maximum observed time in the study is 5.91 years. The Kaplan Meier curve estimates the probability of survival for the 165 individuals with LDL below 100 mg/dL, the 453 individuals with LDL between 100 mg/dL and 159 mg/dL, and the 107 individuals with LDL of at least 160 mg/dL. We can see that, for all observed times, the highest probability of survival belongs to the group with the highest LDL (≥160 mg/dL). By the end of the study, we see that the group with the lowest LDL (<100 mg/dL) has the lowest survival probability, although its survival probability did not start to differ from that of the medium LDL group (100-159 mg/dL) until after about 900 days.

From Cox proportional Hazards regression with a Hubert-White estimator of standard errors, we find that for each 1 mg/dL difference in serum LDL, the risk of death is 0.74% lower in the group with higher LDL. A Wald-based 95% CI suggests that this estimate would not be unusual if the true risk of death in the group with higher LDL were between 0.018% lower and 1.29% lower. We note that 0 is not contained in this interval. This estimate is statistically significant (two-sided p-value=0.009) at the 0.05 level so we have evidence to reject the null hypothesis that the instantaneous risk of death is the same for all observed levels of LDL.

The answer is correct and concise. The descriptive statistics need more detail however (-1). For example, choosing three subgroups may have been underestimating the variance within groups.

9/10

**2.**

**Methods:** See problem 1 for descriptive statistics and survival distributions using Kaplan-Meier estimates across 3 strata based on serum LDL level (<100 mg/dL, 100-159 mg/dL, and ≥160 mg/dL). We also compute a hazard ratio and accompanying Wald-based 95% CI using Cox proportional hazard regression with a Hubert-White sandwich estimator of corresponding standard errors. In view of the possibility that LDL has a multiplicative effect on risk of death, we use a log transformation of continuous serum LDL as the explanatory variable in our proportional hazards regression. Our response variable is an indicator of all-cause mortality.

**Inference:** From Cox proportional hazards regression with a Hubert-White estimator of standard errors, we find that for each twofold increase in serum LDL, the risk of death is 43.6% lower in the group with higher LDL. A Wald-based 95% CI suggests that this estimate would not be unusual if the true risk of death in the group with higher LDL were between 26.2% lower and 56.9% lower. We note that 0 is not contained in this interval. This estimate is statistically significant (two-sided p-value <0.001) at the 0.05 level, so we have some evidence of an association between LDL and instantaneous risk of death under this model.

A two-fold increase in LDL may not be a realistic factor to consider for reporting (-1). Otherwise, the analysis and conclusion are correct.

9/10

**3.**

**Methods:** See problem 1 for descriptive statistics and survival distributions using Kaplan-Meier estimates across 3 strata based on serum LDL level (<100 mg/dL, 100-159 mg/dL, and ≥160 mg/dL). Here, we estimate hazard ratios and accompanying Wald-based 95% CIs using a Cox proportional hazards regression with a quadratic in serum LDL as our explanatory variables. We use the Hubert-White sandwich estimator for corresponding standard errors. Our response variable is an indicator of all-cause mortality.

**Inference:** From Cox proportional Hazards regression, we find that for each 1 mg/dL difference in squared LDL, the risk of death is 0.008% lower in the group with higher LDL. We would like to use this point estimate to detect a non-linear association between LDL and mortality risk, suggesting that a linear model as used in problem 1 may not have been sufficient. A Wald-based 95% CI suggests that the coefficient estimate of the squared term would not be unusual if the true value were between $1.63×10^{-6}$ mg/dL lower and $1.54×10^{-4}$ mg/dL higher. Note that this 95% CI contains 0. The accompanying two-sided p-value of this estimate is 0.055 > 0.05, so we do not have sufficient evidence to conclude that the association between serum LDL and instantaneous risk of death is not linear, where our alternative model is that the association is quadratic.

It is not clear if the linear term was included in the model (-2). A test still needed to be done on the association between mortality and serum LDL (-2).

6/10

**4.**



Figure 2: Comparison of fitted hazard ratios from problems 1-3

As shown in Figure 2, it appears that the fitted hazard ratios from problem 1 (with continuous LDL as the sole explanatory variable) are similar to those from problem 2 (with log-transformed LDL as the sole explanatory variable). Modeling an additive instead of a multiplicative association does not yield substantially different fitted values. The fitted hazard ratio from problem 3 (with a quadratic in LDL as the explanatory variables) seems to have a considerably steeper slope for levels of LDL below 75 mg/dL than the other two specifications. We also observe a slightly steeper slope for values of LDL above 200 mg/dL. The fitted values for all three specifications seem to be approximately the same for LDL between 100 mg/dL and 200 mg/dL, where we have 75% of the observations in our data.

Values were correctly generated for the plot, and observations are correct.

10/10