1.

a.

In the logistic regression model used to answer this question, three distinct groups (those who are of the race black, Asian, other) which each are modeled with a slope, plus an intercept, and it is a saturated model.

Methods:

The odds of subjects having diabetes were compared between subjects who differed by race using whites as a reference group using a logistic regression model. Statistical inference on the ratio of odds of death as a function of serum LDL modeled as a linear continuous variable was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using maximum likelihood. Two-sided p value and 95% confidence interval were computed using the approximate normal distribution for logistic regression parameter estimates.

Results:

735 subjects had data on both race and diabetes presence. Overall, 79 of the 735 had diabetes (10.748%). After correcting for race being black, Asian, or other (comparing the odds of diabetes when examining groups similar in respect to their race), we fail to find a statistically significant overall association between diabetes and race with a p value being .0956 . When examining our reference group (white) (and ignoring issues with multiple comparison), we find that there is a statistically significant association between race and diabetes after controlling for race being black, Asian, or other. Those odds ratio between those who are white vs those who are not is .1085 with a p value of less than .0005. Based on a 95% confidence interval, this result would be unsurprising if the true odds ratio was between .0823563 and .1430144.

b.

The regression parameters for the constant (the reference group white) is explained in part a. For this model, we are comparing the odds ratio between each racial group and the reference, which is whites. Thus, the intercept for each variable is -2.22 (p value = .000), which, when exponentiated, is the odds ratio of having diabetes between those who are white and those who are not (explained in part a). For blacks, the slope is .65678, which corresponds to an odds ratio of having diabetes between those who are black and those who are not of 1.929 (p value = .026). For Asians, the slope is -.4648, which corresponds to an odds ratio of having diabetes between those who are Asian and those who are not of .6282 (p value = .449). For other race, the slope is .6113, which corresponds to an odds ratio of having diabetes between those who are of other race and those who are not of 1.843 (p value = .438).

c.

With .05 level of significance and ignoring multiple comparison issues, I would have found there to be a statistically significant association between being black and diabetes.

d. The report on the overall association between diabetes and race does not change (p value = .0956). This regression model is very similar to part a, but the reference group is now black, thus changing the intercept. Since the Asian and other race variables are being compared to black now, their regression parameters change. The p value of the white group in this model is the same as the p value of the black group in the first model, as you are no just reparameterizing this single comparison.

e.

For this model, we are comparing the odds ratio between each racial group and the reference, which is blacks. Thus, the intercept for each variable is -1.564 (p value = .000), which, when exponentiated, is the odds ratio of having diabetes between those who are black and those who are not of .2093. For whites, the slope is -.6568, which corresponds to an odds ratio of having diabetes between those who are white and those who are not of .5185 (p value = .026). For Asians, the slope is -1.122, which corresponds to an odds ratio of having diabetes between those who are Asian and those who are not of .3258 (p value = .085). For other race, the slope is -.0456, which corresponds to an odds ratio of having diabetes between those who are of other race and those who are not of .9556 (p value = .956).

f.

With .05 level of significance and ignoring multiple comparison issues, I would have found there to be a statistically significant association between being white and diabetes.

g.

If you drop variables that look non significant, you do not come up with the same conclusions when reparameterizing this model, which is wrong.

2.

a.

Descriptive statistics: The study consisted of 735 subjects who were followed for death from any cause for a Kaplan-Meier estimated average of 5.33 years (median 5.66 years, range 5.00 to 5.91 years), during which time 133 deaths were observed. Serum LDL measurements at the time of study enrollment were not available on 10 subjects, two of whom were observed to die after 0.189 and 0.657 years of observation, with the remaining subjects still alive after 5.05 to 5.91 years of observation. In the 725 subjects with available serum LDL measurements at enrollment, the mean LDL was 126 mg/dL (SD 33.6 mg/dL, range 11 to 247 mg/dL).

Table 1 presents estimates of the survival distribution within strata defined by serum LDL and in the combined sample from the 725 subjects with available LDL measurements. The greatest difference in survival distributions is apparent when comparing those individuals having the lowest serum LDL levels (less than 70 mg/dL) at times after 2 years of follow-up. The 5 year survival probability is lowest in that group (59.1%) and is observed highest in the subjects having serum LDL between 160 and 189 mg/dL inclusive (88.0%). On average, the subjects in the lowest LDL stratum were estimated to average 4.91 years of life during the first 5.75 years following study enrollment, while the other strata averaged from 5.23 to 5.45 years. 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 based estimates of distribution of time from study enrollment to death from any cause for subjects having serum LDL measurements at baseline.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Serum LDL at Study Enrollment | | | | | | All Subjects (with LDL Available3) |
|  | 11 – 69 mg/dL | 70 – 99 mg/dL | 100 – 129 mg/dL | 130 – 159 mg/dL | 160 – 189 mg/dL | 190 – 247 mg/dL |
| N Subjects | 22 | 143 | 228 | 225 | 83 | 24 | 725 |
| N Deaths | 10 | 28 | 44 | 34 | 11 | 4 | 131 |
| 2 year Survival Probability \*1 | 100% | 95.8% | 93.9% | 95.6% | 98.8% | 95.8% | 95.6% |
| 5 Year Survival Probability1 | 59.1% | 83.2% | 81.1% | 87.1% | 88.0% | 83.3% | 83.6% |
| 10th Pctile of Survival \*1 | 3.46 y | 3.80 y | 3.41 y | 4.30 y | 4.53 y | 4.13 y | 3.66 y |
| 20th Pctile of Survival \*1 | 3.55 y | 5.44 y | 5.36 y | NA1 | NA1 | NA1 | 5.54 y |
| 5.75 Year Restricted Mean of Survival \*2 | 4.91 y | 5.24 y | 5.23 y | 5.35 y | 5.45 y | 5.32 y | 5.29 y |

\*1 Based on Kaplan-Meier estimates computed within strata defined by LDL and overall. NA indicates that the corresponding percentile is not estimable with the available data.

\*2 Average number of years alive during the first 5.75 years following study enrollment, as computed by the area under Kaplam-Meier survival curves computed within strata defined by LDL and overall

\*3 Ten of the 735 subjects in the study population were missing baseline serum LDL measurements. Two of those subjects were observed to die after 0.189 y and 0.657 years of observation. The remaining 8 subjects with missing LDL data were still alive at the end of their observation period 5.03 to 5.91 years after study enrollment

\*Taken from HW 4 answer key.

**Statistical Methods for inferential statistics:** Distributions of time to death from any cause was compared across groups defined by serum LDL using proportional hazards regression modeling serum LDL when fit as dummy variables using the categories for LDL suggested by the Mayo Clinic. Quantification of association between all cause mortality was summarized by the hazards ratio computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.

**Inferential results:** Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis, we get a p value of .0087, indicating that we can reject the null hypothesis that there is no overall association between LDL and all-cause mortality.

3/3 for descriptive statistics

3/3 for performing an appropriate analysis

1/4 for reporting the association appropriately

Wrong interpretation of coefficient (-2)

Wrong interpretation of CI (-1)

Total: 7

b.

The variables in the model are defined by the dummy variables. For those in the 70 variable (LDL serum level 70-99), the hazard ratio is .398. For those in the 100 varaible, the HR is .393. For those in the 130 variable, the HR is .294. For those in the 160 variable, the HR is .257. For those in the 190 variable, the HR is .316. These hazard ratios are the hazards of death compared to the baseline Hazard of a the variable with serum LDL from 0 to 69 mg/dL.   
  
Total: 5

c. To examine if a continuous linear term fits the data better than this model, I would include a linear predictor term in the model, then use testparm to see if dropping out my fitter value still results in the linear term having a good fit. This results in finding that the linear term does in fact model the data better.

Did not mention what kind of test you use (-1)  
No p-value (-1)  
Wrong conclusion (-1)  
Total: 2

d.

3.

a. See #2 for descriptive statistics.

**Statistical Methods for inferential statistics:** Distributions of time to death from any cause was compared across groups defined by serum LDL using proportional hazards regression modeling serum LDL when fit as linear splines using the categories for LDL suggested by the Mayo Clinic. Quantification of association between all cause mortality was summarized by the hazards ratio computed from the regression model, with confidence intervals and two-sided p values computed using Wald statistics based on the Huber-White sandwich estimator. Subjects missing data for serum LDL at the time of study accrual were omitted from the analysis.

**Inferential results:** Data was available on 725 subjects having mean serum LDL of 126 mg/dL (SD 33.6 mg/dL; range 11 – 247 mg/dL). During an average of 5.33 years of observation, 131 of those subjects were observed to die. From a proportional hazards regression analysis, we get a p value of .0285, indicating that we can reject the null hypothesis that there is no overall association between LDL and all-cause mortality.

b.

For those in the ldl0 group, the hazard ratio is .978 (p value = .019). For those in the ldl70 group, the HR is .980 (p value = .0139). For those in the ldl100 group, the HR is .998 (p value = .835). For those in the ldl130 group, the p value is 1.0036 (p value = .773). For those in the ldl160 group, the HR is .971 (p value = .181). For those in the ldl190 group, the HR is 1.029 (p value = .261).

c.

We can compare the RMSE between the splinefit variable model and the continuous linear fit model. The linear fit model has a lower RMSE and is thus the better fit.

d.

4.

a.

The main advantages the regression strategies used in HWs 4 and 5 have over those used in HWs 1, 2, and 3 are that, in HWs 4 and 5, we do not lose any data by dichotomizing variables and we examine the associations using regression data throughout the range of the variables. This enables us to identify different associations in different ranges of the variables, gaining a deeper understanding of the true association.

b.The figure below shows all of the Hazard Ratios modeled on HW 4 and 5. They are overall very similar, but those from HW 5 on average have lower HRs and have more linear of a fit. Also, Those from HW 5 do not change as much at lower LDL levels.



c.

A priori, I would have done a proportional hazard regression using the logarithmic transformed hazard ratio in order to get a clear understanding for the presumably linear association between log LDL and hazard. I would not have had any indication that the relationship was nonlinear or be different along different ranges of the variables, so I would not have opted for more complex dummy or spline based modeling.