HW 5

1a. This is a saturated model.

**Methods**: Robust logistic regression was used to find an association between prevalence of diabetes and race (i.e. odds of diabetes diagnosis across race groups). Race was categorized as white, black, Asian, and other with whites being used as the reference group. Diabetes was used as a binary variable (yes/no). Statistical inference was based on Wald statistics with 95% confidence intervals and two sided p-values computed using the approximate normal distribution for logistic regression parameter estimates.

**Results**: The difference between race groups relative to the odds of diabetes diagnosis is not different from what we would expect if race had no true effect (p=0.0956). Blacks had an odds of diabetes diagnosis 1.92 times as large as whites (95% CI: 1.08, 3.44), Asians had an odds of diabetes diagnosis 0.63 times that of whites (95% CI: 0.19, 2.09), and those in the other race group had an odds of diabetes diagnosis 1.84 times as large as whites (95% CI: 0.39, 8.63).

1b. The odds of diabetes diagnosis among whites is 0.11. The odds of diabetes diagnosis among blacks is 1.93 times that of whites. The odds of diabetes diagnosis among Asians is 0.63 times that of whites. The odds of diabetes diagnosis among those in the other race group is 1.84 times that of whites.

1c. The odds of diabetes diagnosis among whites is statistically different from zero (p<0.001). Diabetes diagnosis among blacks is significantly different from that of whites (p=0.026). Diabetes diagnosis among either Asians or those in the other race group is not statistically different from whites (p=0.449 and 0.438 respectively).

1d. The overall association between race and diabetes would be the same as in part a. However, the odds ratios for each race group would be relative to blacks instead of whites. The intercept would be the odds of diabetes among blacks instead of among whites.

1e. The odds of diabetes diagnosis among blacks is 0.21. The odds of diabetes diagnosis among whites is 0.52 times that of blacks. The odds of diabetes diagnosis among Asians is 0.33 times that of blacks. The odds of diabetes diagnosis among those in the other race group is 0.96 times that of blacks.

1f. The odds of diabetes diagnosis among blacks is statistically different from zero (p<0.001). The odds of diabetes diagnosis among whites is statistically different from that of blacks (p=0.026). The odds of diabetes diagnosis among Asians and those in the other race groups are not statistically different from that of blacks (p=0.085 and 0.956 respectively).

1g. In parts c and f, the difference in the odds of diabetes diagnosis was different between blacks and whites only, while the overall association between race and diabetes diagnosis was insignificant. Stepwise model building might include race into the model just because whites and blacks had a significant p-value, even though race as a whole was not associated with diabetes diagnosis.

2a. **Descriptive statistics:**

**Methods**: Kaplan Meier estimates of 10, 50, and 90th percentiles, and mean time of follow-up were calculated for the censoring distribution, as well as the minimum and maximum observed censored times. Serum LDL was categorized as suggested by the Mayo Clinic: 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. The number of subjects and deaths within each group was calculated. Kaplan Meier estimates were graphed, as well as calculations of 2 and 5-year survival probabilities, 10th and 20th percentile of the survival distribution, and restricted mean survival for 5.75 years (time in which all strata of LDL still had some subjects at risk).

**Results**: There were 735 total subjects in this study (followed for a Kaplan Meier estimate restricted mean of 5.33 years, range 5-5.91 years), and 133 died from any cause. 10 subjects did not have serum LDL measurements, of which 2 died. Mean LDL was 126 mg/dL (SD: 33.6 mg/dL, range 11-247 mg/dL). The lowest LDL group (11-69 mg/dL) had the highest 2-year survival probability but the lowest 5-year survival probability. The 11-69 mg/dL group also had the lowest survival for the first 5.75 years, as well as 20th percentile survival. From the KM graph, the lowest LDL group diverged in survival from the other groups after about 3 years.

Table: Kaplan Meier estimates by LDL group

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **11-69 mg/dL** | **70-99 mg/dL** | **100-129 mg/dL** | **130-159 mg/dL** | **160-189 mg/dL** | **190-247 mg/dL** | **All subjects** |
| Subjects | 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 | 95.6 |
| 5 year survival probability (%) | 59.1 | 83.2 | 81.1 | 87.1 | 88.0 | 83.3 | 83.6 |
| 10th percentile survival (y) | 3.46 | 3.80 | 3.41 | 4.30 | 4.53 | 4.13 | 3.66 |
| 20th percentile survival (y) | 3.55 | 5.44 | 5.36 | NA | NA | NA | 5.54 |
| 5.57 year restricted mean of survival (y) | 4.91 | 5.24 | 5.23 | 5.35 | 5.45 | 5.32 | 5.29 |

NA: unable to calculated from data.



**Inferential statistics:**

**Methods**: Time to death from any cause was relative across the following LDL groups: 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. Hazard ratios from cox regression model were used, with confidence intervals and two-sided p-values calculated with Wald statistics based on the Huber-White sandwich estimator.

**Results**: For the 725 subjects with LDL measurements, the mean LDL was 126 mg/dL (SD: 33.6 mg/dL, range 11-247 mg/dL). 131 subjects died during the time of analysis. A two-sided p-value of 0.0087 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is not associated with LDL levels, with the tendency of higher mortality for those with lower LDL levels. The hazard for the 70-99 mg/dL group is 0.40 times that of the 11-69 mg/dL group (95% CI: 0.20-0.78). The hazard for the 100-129 mg/dL group is 0.39 times that of the 11-69 mg/dL group (95% CI: 0.21, 0.74). The hazard of the 130-159 mg/dL group is 0.29 times that of the 11-69 mg/dL group (95% CI: 0.15, 0.57). The hazard of the 160-189 mg/dL group is 0.26 times that of the 11-69 mg/dL group (95% CI: 0.11-0.58). The hazard of the 190-247 mg/dL group is 0.32 times that of the 11-69 mg/dL group (95% CI: 0.10, 0.99).

2b. The estimated risk of death is 60% lower for those in the 70-99 mg/dL group relative to those in the 11-69 mg/dL group.

The estimated risk of death is 61% lower for those in the 100-129 mg/dL group relative to those in the 11-69 mg/dL group.

The estimated risk of death is 71% lower for those in the 130-159 mg/dL group relative to those in the 11-69 mg/dL group.

The estimated risk of death is 74% lower for those in the 160-189 mg/dL group relative to those in the 11-69 mg/dL group.

The estimated risk of death is 68% lower for those in the 190-247 mg/dL group relative to those in the 11-69 mg/dL group.

The relative risk of death is 1 for those in the 11-69 mg/dL group.

2c. The log likelihood ratio test would be appropriate. From the LR test, the current categorized LDL model is not a significantly better fit than the model using only a continuous linear term for LDL (p=0.4776).

2d. See part 4

3a. **Descriptive statistics**: see 2a.

**Inferential statistics:**

**Methods**: Linear splines were created with the following “knots”: 70, 100, 130, 160, 190. Hazard ratios from cox regression model were used, with confidence intervals and two-sided p-values calculated with Wald statistics based on the Huber-White sandwich estimator.

**Results**: For the 725 subjects with LDL measurements, the mean LDL was 126 mg/dL (SD: 33.6 mg/dL, range 11-247 mg/dL). 131 subjects died during the time of analysis. A two-sided p-value<0.0001 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is not associated with LDL levels.

There is a 2.2% decrease (95% CI: 0%, 4.0%) in risk of death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 11-69 mg/dL group. There is a 2.0% decrease (95% CI: 4.7% decrease, 0.7% increase) in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 70-99 mg/dL group. There is a 0.2% decrease (95% CI: 2.4% decrease, 1.9% increase) in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 100-129 mg/dL group. There is a 0.4% increase (95% CI: 2.1% decrease, 2.8% increase) in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 130-159 mg/dL group. There is a 3.1% decrease (95% CI: 7.0% decrease, 1.4% increase) in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 160-189 mg/dL group. There is a 3.9% increase (95% CI: 2.1% decrease, 8.1% increase) in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 190-247 mg/dL group.

3b. There is a 2.2% decrease in risk of death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 11-69 mg/dL group.

There is a 2.0% decrease in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 70-99 mg/dL group.

There is a 0.2% decrease in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 100-129 mg/dL group.

There is a 0.4% increase in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 130-159 mg/dL group.

There is a 3.1% decrease in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 160-189 mg/dL group.

There is a 3.9% increase in death corresponding to an increase in LDL of 1 mg/dL between two groups both in the 190-247 mg/dL group.

3c. The log likelihood ratio test would be appropriate. From the LR test, the current categorized LDL model is not a significantly better fit than the model using only a continuous linear term for LDL (p=0.3730).

3d. See part 4

4a. Homework 4 and 5 used nonlinear models. The main advantage of this would be the ability to model nonlinear associations more accurately. A nonlinear association may not be detected using the approaches in Homework 1-3. Centering variables (i.e. ldl) can reduce collinearities as well.

4b.



The overall trend using dummy variables (blue) and splines (red) is similar to that from Homework 4 (i.e. generally downward trend).



The similarities are seen above with all five fitted models. The main difference is in the dummy model in the lowest LDL group, which is due to the higher variation of hazard.

4c. Although the association between all cause mortality and serum LDL in the elderly population is not exactly linear, a simple linear regression model would probably be the best analysis since the other models are relatively close to the linear regression model. The quadratic model would be harder to interpret while the spline may lead to loss of precision from loss of degrees of freedom. The dummy variable model does not give a particularly good fit at lower LDL levels. Simple linear regression would be the easiest to understand.