BIOS 515

Hw 5

1. (a) This model is saturated since there are four parameter estimates that correspond to the four race factors with white being the reference factor.

**Methods**

To evaluate the association between prevalence of diabetes and race, I proceed by comparing the odds of diabetes diagnosis across different races with the whites as the reference group. It’s not specified in the problem which logistic regression to use and I proceed with robust logistic regression. Statistical inference is based on the Wald statistic computed from the regression slope parameter and its standard error, with two –sided p value and 95% confidence interval computed using the approximate normal distribution for logistic regression parameter estimates.

**Inference**

The study sample consisted of 735 subjects whose diabetes records were present. This included 572 whites, 86 black, 44 Asians, and 10 subjects belonging to other race. There were 56 whites, 18 blacks, 3 Asians, and 2 other subjects who were diagnosed with diabetes.

The likelihood ratio chi-square of 6.35 with p-value of 0.0956 tells that the full model is not a better fit compared to a model with no predictors. We therefore fail to reject with high confidence the null hypothesis that diabetes is associated with race.

**(b)** **Interpretation of regression parameters**

The odds of a black subject being diagnosed with diabetes compared to being a white subject were 92.86% higher (odds ratio 1.9286). The 95% CI tells us that the observed odds ratio would not be unusual if the true value were anywhere between (1.082, 3.439).

The odds of an Asian being diagnosed with diabetes compared to whites were 37.18% lower (odds ratio 0.6282). The 95% CI tells that this observed in odds would not be unusual if the true odds ratio were anywhere between (0.1888, 2.091).

The odds of other races being diagnosed with diabetes compared to whites were 84.29% higher (odds ratio 1.8429). The 95% CI tells that this observed in odds would not be unusual if the true odds ratio were anywhere between (0.3935, 8.631).

The odds of other races being diagnosed with diabetes compared to blacks were 84.29% higher (odds ratio 1.8429). The 95% CI tells that this observed in odds would not be unusual if the true odds ratio were anywhere between (0.3935, 8.631).

The intercept tells us that the odds of a white subject being diagnosed with diabetes were 89.15% lower (odds ratio 0.1085). The 95% CI tells that this observed in odds would not be unusual if the true odds were anywhere between (0.08235, 0.1430).

(c) If we were to ignore issue related to multiple comparisons, at an alpha level of 0.05, I would remove the Asians (p-value 0.449) and other (p-value 0.438) races from the regression model since we would obtain a better fit without them. The p-values for blacks (p-value 0.026) and whites (p<0.0001) are significant and fitting a model with these two races as the predictors of interest would provide a better fit for the model.

(d)

There is no change in the overall formal inference from the one provided in part since it’s a re-parameterization (coded differently) of the model in a. Compared to the regression model in part a. The change in reference group from white to black however result to different odds ratios within the respective races.

(e)

The odds of a white subject being diagnosed with diabetes compared to a black subject were 48.15% lower (odds ratio 0.5185). The 95% CI tells us that the observed odds ratio would not be unusual if the true value were anywhere between (0.2908, 0.9246).

The odds of an Asian subject being diagnosed with diabetes compared to a black subject were 67.42% lower (odds ratio 0.3258). The 95% CI tells us that the observed odds ratio would not be unusual if the true value were anywhere between (0.0909, 1.167).

The odds of other races being diagnosed with diabetes compared to blacks were 4.44% lower (odds ratio 0.9556). The 95% CI tells that this observed in odds would not be unusual if the true odds ratio were anywhere between (0.1925, 4.742).

The intercept tells us that the odds of a black subject being diagnosed with diabetes were 79.07% lower (odds ratio 0.2093). The 95% CI tells that this observed in odds would not be unusual if the true odds were anywhere between (0.1259, 0 .3480).

(f)

Ignoring issue related to multiple comparisons, only white and black would be included in the regression model due to their significant p-values. Asian (p-value 0.085) and other (p-value 0.956) races are non-significant and at alpha level of 0.05, they won’t be included in the model.

(g) The use of p-values for individual regression parameters from dummy variable regression to decide whether to include or exclude those parameters in the regression model can be misleading. In this case, race as a factor is insignificant but the p-values from the dummy variables corresponding to black and white are significant.

It would be more appropriate to include race in the regression model to avoid any loss of information that can arise from variable selection using p-values for individual regression parameters from dummy variable regression.

1. (a) **Methods**

The descriptive statistics for censoring distribution included the minimum and maximum observed censoring times and the Kaplan-Meier estimates of the 10th, 50th, and 90th percentiles. The probability of survival from the Kaplan-Meier estimate of the censoring distribution’s survivor curve is also provided.

For serum LDL levels, the number of cases with missing data, the minimum, maximum, mean, standard deviation, and the 25th, 50th, and 75th percentiles were included. Serum LDL was categorized into 8 categories as suggested by Mayo Clinic: less than 70 mg/dl, 70-99 mg/dl, 120-129 mg/dl, 130-159mg/dl, 160-189 mg/dl, and greater than 190 mg/dl. The Kaplan-Meier estimates of survival were calculated and graphed and estimates of survival probabilities for 2 and 5 year given.

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 Kaplan-Meir estimates**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Serum LDL at study Enrollment** | | | | | | **All subjects (with LDL available)** |
|  | **11-69 mg/dl** | **70-79 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** | **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%** | **83.3%** | **83.6%** |
| **10th pctile of survival** | **3.46y** | **3.80y** | **3.41y** | **4.39y** | **4.53y** | **4.13y** | **3.66y** |
| **20th pctile of surivival** | **4.91y** | **5.44y** | **5.36y** | **NA** | **NA** | **NA** | **5.54y** |



**Inference**

Distributions of time to death from any cause was compared across groups defined by serum LDL at baseline using proportional hazards regression modeling serum LDL as a continuous untransformed random variable. 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.

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 estimate that the instantaneous risk of death is a relative 7.14% lower (hazard ratio 0.929) for each 10 mg/dL higher serum LDL level at baseline. Based on a 95% confidence interval, this observed hazard ratio suggesting lower death rates for groups of patients with higher LDL levels would not be judged unusual if the true instantaneous risk of death were anywhere from 1.80% to 12.2% lower in a group having baseline serum LDL 10 mg/dL higher than that in another group (95% CI for hazard ratio 0.878 to 0.982). A two-sided p value of 0.009 suggests that we can with high confidence reject the null hypothesis that the risk of death from any cause is not associated with serum LDL levels in favor of a tendency for lower mortality with higher serum LDL levels.

(b)

The intercept represents the reference group, (0, 70) mg/dl and its hazard ratio is of no clinical importance since it compares it with itself.

By looking at hazard ratios for different groups, we observe that subjects who had LDL levels between 70 and 100 mg/dl had about 60.2% (hazard ratio 0.3980) less risk of death for each unit increase in LDL compared to those subjects with LDL levels lower than 70mg/dl (reference group). The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.2026, 0.7820).

Subjects with LDL levels between 100 and 130 mg/dl had about 60.74% (hazard ratio 0.3926) less risk of death for each unit increase in LDL compared to those subjects with LDL levels lower than 70mg/dl (reference group). The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.2071, 0.7442).

Subjects with LDL levels between 130 and 160 mg/dl had about 70.61% (hazard ratio 0.2939) less risk of death for each unit increase in LDL compared to those subjects with LDL levels lower than 70mg/dl (reference group). The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.1521, 0.5678).

Subjects with LDL levels between 160 and 190 mg/dl had about 74.35% (hazard ratio 0.2565) less risk of death for each unit increase in LDL compared to those subjects with LDL levels lower than 70mg/dl (reference group). The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.1135 0.5799).

Subjects with LDL levels above 190 mg/dl had about 68.33% (hazard ratio 0.3167) less risk of death for each unit increase in LDL compared to those subjects with LDL levels lower than 70mg/dl (reference group). The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.1014, 0.9892).

(c)

To assess if this regression model provides a better fit than does a model that uses only a continuous linear term for LDL, we perform a test of linearity (all dummy variables equal 0) using an chi-square test.

Based on the chi-square p-value (0.3988), we can’t prove there is no non-linearity although we can’t claim linearity either.

(d)

See the plot in 4 (b)

1. (a)

See part 2 (a)

(b)

Subjects with LDL levels between 0 and 70 mg/dl had about 2.19% (hazard ratio 0.9781) less risk of death for each unit increase in LDL. The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.9602, 0.9963).

Subjects with LDL levels between 70 and 100 mg/dl had about 2.03% (hazard ratio 0.9797) less risk of death for each unit increase in LDL. The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.9535, 1.007).

Subjects with LDL levels between 100 and 130 mg/dl had about 0.23% (hazard ratio 0.9977) less risk of death for each unit increase in LDL. The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.9764, 1.019).

Subjects with LDL levels between 130 and 160 mg/dl had about 0.36% (hazard ratio 0.9977) higher risk of death for each unit increase in LDL. The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.9794 1.0284).

Subjects with LDL levels between 160 and 190 mg/dl had about 2.91% (hazard ratio 0.9709) less risk of death for each unit increase in LDL. The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.9298 1.014).

Subjects with LDL levels above 190 mg/dl had about 2.88% (hazard ratio 1.029) higher risk of death for each unit increase in LDL. The 95% CI suggests that this observation won’t be unusual if the true hazard ratio was anywhere between (0.9791, 1.081).

(c)

To assess if this regression model provides a better fit than does a model that uses only a continuous linear term for LDL, we perform a test of linearity using chi-square test.

Based on the observed p-value (0.0788), we can’t prove there is no non-linearity although we can’t claim linearity either.

(d)

See the plot in 4(b)

1. (a)

The regression strategies used in homework 4 and 5 address censoring as opposed to the techniques used in home works 1-3 where we used linear, logistic, and Poisson regression. Use of hazard regression in home works 4 and 5 ensures that the censored subjects are addressed hence provides us with more information.

In addition regression strategies for home works 4 and 5 allow for possibility of non-linear trends which allows for more information capturing with less forcing in the fit.

**(b) Similarities**

Below is graph for the fitted values of hazards from problem 2 and 3 parts d



The fitted values of models in both home works 4 and 5 demonstrate a downward trend and are centered at 160 mg/dl

**Differences**

The fitted values from homework 4 are smooth while those in this homework are not. The fitted values for the LDL groups from problem two demonstrate some jumps since they are categorical. The splines appear a little smooth though not anything close to those from homework 4.

(c)

I would prefer using splines as a priori for exploring an (unadjusted) association between all-cause mortality and serum LDL in an elderly population because it provides more flexibility and less borrowing of information. Furthermore splines don’t over-specify as in quadratic and linear modelling but still tests linearity.