**Biostats 518 - Homework 3**

**Question 1**

1. Yes. This is a saturated model because there are two groups (those who have high LDL (≥160 mg/dL) and those who do not have high LDL), modeled with two regression parameters (slope and intercept).
2. Among subjects with low LDL, the estimated odds of dying is 0.205. From this odds, the estimated probability of dying within 5 years is 0.196. The observed proportion who died within 5 years is 0.1699. The estimate odds of dying within 5 years, and the estimated probability of dying within 5 years are both greater than the observed proportion of subjects who died within 5 years (by a difference of 0.0351 and .0261, respectively), among subjects with low LDL.
3. Among subjects with high LDL, the estimated odds of dying is 0.151. From this odds, the estimated probability of dying within 5 years is 0.131. The observed proportion who died within 5 years is 0.131. The estimate odds of dying within 5 years is 0.02 greater than the observed proportion who died, and the estimated probability is approximately equal to the observed proportion of subjects who died within 5 years, among subjects with high LDL.
4. Using logistic regression with robust variance, the odds of death within 5 years of study enrollment were compared between subjects with and without high serum LDL (≥160 mg/dL). An odds ratio was different from 1 was tested and 95% confidence intervals were computed using robust standard errors. Of the 618 subjects without high LDL, the odds of dying within 5 years was 0.205, and for the 107 subjects with high LDL, the odds of dying within 5 years was 0.151. Based on 95% confidence intervals, the observed odds ratio (OR) of 0.735 would comparing those with high LDL to those with low LDL would not be unusual if the true OR was between 0.40 and 1.340. The two-sided p-value of 0.316 suggests that there is insufficient evidence to reject the null hypothesis that the odds of dying within 5 years are not associated with serum LDL levels.

 In Question 6 on homework 1, which used Fisher’s exact test, the point estimates for the odds and odds ratio were the same. However, the 95% confidence interval was wider (0.373 to 1.36.), and the p-value was larger (0.396) when using Fisher’s exact test,

 compared to logistic regression. This is to be expected because the standard error is slightly smaller when using robust standard error in the logistic regression. Additionally when calculating the 95% CIs in the logistic regression, the critical value is based on 723 degrees of freedom whereas 158.7 degrees of freedom are used for Fisher’s exact test. The larger the critical value, the wider the 95% CIs.

 However, in both methods, the ultimate conclusion remains the same: given the p-value

 greater than 0.05, there is insufficient evidence to reject the null hypothesis that the

 odds of dying within 5 years are not associated with serum LDL levels.

1. If we instead fit a regression model for 5 year mortality with low LDL as the predictor (instead of high LDL), the odds of death in both groups remains the same. The p-value and standard error also remain the same. However, the odds ratio (and the 95% CIs) is the reciprocal, using high LDL as the reference group instead of low LDL. The OR comparing those with low LDL to those with high LDL is 1.360 (95% CI: 0.746, 2.478), which is the reciprocal of the OR in part D above. Therefore, the overall interpretation remains the same.

If we instead fit a regression model for 5-year survival (instead of mortality) with high LDL as the predictor, the odds of survival is the reciprocal of the odds of death. The odds of 5 year survival among those with low LDL is 4.89 and among those with high LDL is 6.64, which is equal to the reciprocal of the odds reported in parts B and C, respectively. The OR is when using the survival as the outcome also the reciprocal of the OR when using mortality as the outcome, 1.360 (95% CI: 0.746, 2.478), which is the same as when low LDL is used instead of high LDL with mortality at the outcome. The standard error and p-value remain the same, as does the overall interpretation of the relationship between LDL and mortality.

1. If we used 5 year mortality as the predictor, and high LDL as the outcome, the odds ratio, confidence intervals, p-value, and interpretation would remain exactly the same. The OR has this property. The OR of death within 5 years comparing those with high LDL to those with low LDL is equal to the OR of high LDL comparing those who died within 5 years to those who survived. The interpretation remains the same so we fail to reject the null of no association between LDL and 5 year mortality.

**Question 2**

1. Yes. This is a saturated model because there are two groups (those who have high LDL (≥160 mg/dL) and those who do not have high LDL), modeled with two regression parameters (slope and intercept).
2. The estimated probability of dying within 5 years among those with low LDL is 0.1699. This is equal to the observed proportion of subjects with low LDL who died within 5 years. The odds of dying within 5 years among those with low LDL is 0.205, which is 0.04 higher than the observed proportion who died within 5 years.
3. The estimated probability of dying within 5 years among those with low LDL is 0.131. This is equal to the observed proportion of subjects with high LDL who died within 5 years. The odds of dying within 5 years among those with high LDL is 0.151, which is 0.02 higher than the observed proportion who died within 5 years.
4. Using linear regression with robust variance, the probability of death within 5 years of study enrollment were compared between subjects with and without high serum LDL (≥160 mg/dL). Among the 618 subjects with low LDL, the probability of death within 5 years was 0.170, and among the 107 subjects with high LDL, the probability of death within 5 years was 0.131. Based on 95% confidence intervals, the observed 3.91% lower absolute survival probability in subjects with higher serum LDL would not be unusual if the true difference in the probability of death were between 11.0% lower and 3.16% higher in the high LDL group compared to the low LDL group. The two-sided p-value of 0.278 suggests that there is insufficient evidence to reject the null hypothesis that the probability of dying within 5 years is not associated with serum LDL levels. In Question 5 on homework 1, which used Fisher’s exact test, the point estimates for the probabilities of death were the same. However, the 95% confidence interval for the difference was narrower (3.14%-10.9%). However, in both methods, the ultimate conclusion remains the same: given the p-value greater than 0.05, there is insufficient evidence to reject the null hypothesis that the odds of dying within 5 years are not associated with serum LDL levels.
5. If we instead fit a regression model for 5 year mortality with low LDL as the predictor (instead of high LDL), the probability of death in both groups, the absolute difference and absolute 95% CIs, and the p-value and standard error for the difference remain the same. However, the reference category is high LDL instead of low LDL as it was previously.

If we instead fit a regression model for 5-year survival (instead of mortality) with high LDL as the predictor, the difference in survival, 95% CIs, SE, and p-value for would remain the same as the probability of mortality. However, the reference category would instead be the probability of survival among those with low LDL (0.830), which is 1 minus the probability of death among those with low LDL as stated previously (0.170).

1. The probability of high LDL is 0.153 among those who survived at least 5 years and 0.118 among those who died within 5 years. The difference in the probability of having high LDL is -0.0358 (95% CI: -0.101, 0.0290). While the estimates are different in their meaning and interpretation, the p-value remains the same (0.278), so the interpretation will also remain the same and we fail to reject the null of no association between LDL and 5 year mortality.

**Question 3**

1. Yes. This is a saturated model because there are two groups (those who have high LDL (≥160 mg/dL) and those who do not have high LDL), modeled with two regression parameters (slope and intercept).
2. The estimated probability of dying within 5 years among those with low LDL is 0.1699. This is equal to the observed proportion of subjects with low LDL who died within 5 years. The odds of dying within 5 years among those with low LDL is 0.205, which is 0.04 higher than the observed proportion who died within 5 years.
3. The estimated probability of dying within 5 years among those with low LDL is 0.131. This is equal to the observed proportion of subjects with high LDL who died within 5 years. The odds of dying within 5 years among those with high LDL is 0.151, which is 0.02 higher than the observed proportion who died within 5 years.
4. Using poisson regression with robust variance, the probability of death within 5 years of study enrollment were compared between subjects with and without high serum LDL (≥160 mg/dL). Among the 618 subjects with low LDL, the probability of death within 5 years was 0.170, and among the 107 subjects with high LDL, the probability of death within 5 years was 0.131. Based on 95% confidence intervals, the observed risk ratio of 0.77 comparing those with high LDL to those with low LDL would not be unusal if the true risk ratio were between 0.458 and 1.29. The two-sided p-value of 0.324 suggests that there is insufficient evidence to reject the null hypothesis that the probability of dying within 5 years is not associated with serum LDL levels.

 In Question 5 on homework 1, which used Fisher’s exact test, the point estimates for the probabilities were the same. The risk ratio from this method resulted in the same point value of 0.77 and with 95% CIs of 0.458-1.29, and are approximately equal to that

 found using poisson regression, due to using robust standard error.

1. If we instead fit a regression model for 5 year mortality with low LDL as the predictor (instead of high LDL), the probability of death in both groups remains the same. The p-value and standard error also remain the same. However, the risk ratio (and the 95% CIs) is the reciprocal, using high LDL as the reference group instead of low LDL. The RR comparing those with low LDL to those with high LDL is 1.30 (95% CI: 0.773, 2.18), which is the reciprocal of the RR in part D above. The overall interpretation remains the same, and the RR is the reciprocal so uses a different reference in the interpretation.

If we instead fit a regression model for 5-year survival (instead of mortality) with high LDL as the predictor, the relative risk of survival is 1.047 (95% CI: 0.965, 1.14) comparing those with high LDL to those who low LDL. This does not retain the same properties or the odds and difference, and therefore provides different information. The SE and the p-value also differ from parts a-d (p=0.270). However, we would still fail to reject the null of no association between LDL and survival based on this two-sided p-value.

1. The RR of high LDL comparing those who died within 5 years to those who survived is 0.767 (95% CI: 0.457, 1.30; p-value=0.323). This is similar to the estimate of the RR of death comparing those with high LDL to those with low LDL (RR=0.770; 95% CI: 0.458, 1.29; p-value=0.324). Thus we should still fail to reject the null of no association between LDL and survival based on this two-sided p-value.

**Question 4**

1. When fitting a linear regression model for 5 year mortality using serum LDL as a continuous predictor variable, we get a difference in mortality probability of -0.00103, compared to those with a 1 mg/dL lower LDL. Based on a 95% confidence interval, this observed difference in probability would not be unusual if the true difference in mortality were between -0.00188 and -0.000185. Based on a two-sided p-value of 0.019, we can reject the null that there is no association between serum LDL and 5-year mortality.
2. When fitting a poisson regression model for 5 year mortality using serum LDL as a continuous predictor variable, we get an risk ratio of 0.994 (95% CI: 0.988, 0.999). Thus, for every 1 mg/dL increase in LDL, the risk of death within 5 years decreased by a factor of 0.006. The risk ratio of death within 5 years is 0.994 comparing those with 1 mg/dL greater LDL to those with 1 mg/dL lower LDL. Based on 95% confidence intervals, this observed RR would not be unusual if the true risk ratio of death was between 0.988 and 0.999. Based on a two-sided p-value of 0.018, we can reject the null that there is no association between serum LDL and the 5-year mortality.
3. When fitting a logistic regression model for 5 year mortality using serum LDL as a continuous predictor variable, we get an odds ratio of 0.992 (95% CI: 0.986, 0.999). Thus, for every 1 mg/dL increase in LDL, the odds of death within 5 years decreased by a factor of 0.008. The odds ratio of death within 5 years is 0.992 comparing those with 1 mg/dL greater LDL to those with 1 mg/dL lower LDL. Based on 95% confidence intervals, this would not be unusual if the true odds ratio of death was between 0.986 and 0.999. Based on a two-sided p-value of 0.019, we can reject the null that there is no association between serum LDL and the 5-year mortality.
4. The conclusion made when treating LDL as a continuous variable (where we do reject the null of no association with 5 year mortality) differs from those made in problems 1-3 with LDL dichotomized, where we failed to reject the null hypothesis of no association.

 Analyses I would prefer a-priori include:

 -Determine if is it more precise to have a dichotomized or continuous measurement.

 -Would a multiplicative model be preferred or an additive model.

 -Determine if the difference in the probability of death between LDL groups is more

 useful for the goals of the study, or if the association between LDL and death is more

 useful.

 -If the proportion of death can be calculated directly, the odds is not necessary to describe

 the relationship. This depends on the study design. It is better to use a more

 interpretable estimate.

 -Should LDL be log transformed? What is the linearity/distribution of LDL?