Biostats 518 HW#3

Problem 1 = 14/25

1a. It is a saturated regression model, since there are two groups (high and low LDL) being modeled with two regression parameters (intercept and slope).

1b. The odds of dying within 5 years for subjects with low LDL is 0.205. Using the formula probability=odds/(1+odds), we get a probability of dying within 5 years to be 0.17. This probability is the same as the observed proportion of subjects with low LDL dying within 5 years. The odds are the observed proportion /(1-observered proportion).

1c. The odds of dying within 5 years for subjects with high LDL is 0.151. The probability of dying within 5 years is 0.177. This probability is the same as the observed proportion of subjects with low LDL dying within 5 years. The odds are the observed proportion /(1-observered proportion).

1d. **Method**: Logistic regression was used to look at the association between 5-year mortality and high LDL levels (>=160mg/dL) in a sample of 735 elderly subjects between the ages of 65 and 99. 5-year survival and LDL were dichotomized. A significance level of 0.05 was used.

**Results**: We estimate that the odds of dying within 5 years is 26.5% lower in those with high LDL compared to those with low LDL, though this estimate is not statistically significant (p=0.32). From the 95% confidence interval, it would not be unusual if the odds in the high LDL group was anywhere from 59.6% lower to 34.0% higher than in the low LDL group.

We rejected the null in both this problem and in homework 1 #5, and the p-values were similar. We got the same odds ratio as in problem 6, although the p-values and confidence intervals are different. The difference is because we dichotomized the variables here, which lead to loss of information, thus increasing the p-value and confidence interval.

1e. In changing the indicator to low LDL, the odds within each group would be the same (low LDL, high LDL), but the odds ratio would be in the inverse (1/OR). The same would occur for changing the predictor to at least 5 years of survival. If you changed both the predictor and response variables, it would remain the same. Either way it is still a saturated model

1f. It would be a saturated model as well. If we looked at the distribution of LDL across vital status groups, then our model would be finding the odds of high LDL for subjects who survived/died within 5 years, which is different from what we currently have for b and c. However, the odds ratio would remain the same.

Problem 2 = 12/25

2a. It is a saturated regression model, since there are two groups (high and low LDL) being modeled with two regression parameters (intercept and slope).

2b. For subjects with low LDL, the estimated probability of dying within 5 years is 0.17. The estimated odds, using the formula prop/(1-prop), are 0.205. The estimated probability is the same as the observed proportion of subjects with low LDL dying within 5 years. The odds are the observed proportion divided by 1 minus the observed proportion.

2c. For subjects with high LDL, the estimated probability of dying within 5 years is 0.13. The estimated odds, using the formula prop/(1-prop), are 0.151. The estimated probability is the same as the observed proportion of subjects with high LDL dying within 5 years. The odds are the same as prop/(1-prop), where the proportion is the observed proportion of subjects with high LDL dying within 5 years.

2d. **Method**: Linear regression was used to look at the association between 5-year mortality and high LDL levels (>=160mg/dL) in a sample of 735 elderly subjects between the ages of 65 and 99. 5-year survival and LDL were dichotomized. A significance level of 0.05 was used.

**Result**: We estimate the probability of dying within 5 years to be 3.91% lower in those with high LDL compared to those with low LDL, although this difference was not significant (p=0.315). From the 95% confidence interval, it would not be unusual if the proportion of dying within 5 years for those with high LDL were between 11.5% lower and 3.7% higher than those with low LDL.

We rejected the null in HW1 #5 and 6 as well, and the difference was the same as in problem 5 although the p-value and confidence intervals are different. The difference is because we dichotomized the variables here, which lead to loss of information, thus increasing the p-value and confidence interval.

2e. The proportion and odds within each group would be the same, but the difference would be the negative (opposite) of a-c.

2f. It would still be a saturated model. If we looked at the distribution of LDL across vital status groups, then our model would be finding the probability of high LDL for subjects who survived/died within 5 years, which is different from what we currently have for b and c. The difference would be different as well.

Problem 3 = 11/25

3a. It is a saturated regression model, since there are two groups (high and low LDL) being modeled with two regression parameters (intercept and slope).

3b. For subjects with low LDL, the estimated probability of dying within 5 years is 0.17. The estimated odds, using the formula prop/(1-prop), are 0.205. The estimated probability is the same as the observed proportion of subjects with low LDL dying within 5 years. The odds are the observed proportion divided by 1 minus the observed proportion.

3c. For subjects with high LDL, the estimated probability of dying within 5 years is 0.13. The estimated odds, using the formula prop/(1-prop), are 0.151. The estimated probability is the same as the observed proportion of subjects with high LDL dying within 5 years. The odds are the same as prop/(1-prop), where the proportion is the observed proportion of subjects with high LDL dying within 5 years.

3d. **Method**: Poisson regression was used to look at the association between 5-year mortality and high LDL levels (>=160mg/dL) in a sample of 735 elderly subjects between the ages of 65 and 99. 5-year survival and LDL were dichotomized. A significance level of 0.05 was used.

**Result**: The rate ratio of dying within 5 years for those with high LDL relative to those with low LDL is 0.77, although it is not statistically significant (p=0.359). From the 95% confidence interval, it would not be unusual if the rate ratio of dying within 5 years for those with high LDL compared to those with low LDL were between 0.44 and 1.35.

We rejected the null in both 3d and in homework 1. However, the point estimate, p-value, and confidence interval are different. It is different because we are looking at the rate ratio, which is modeled differently from linear regression and logistic regression in homework 1. We also have dichotomized variables.

3e. The rate ratio would be the inverse in both cases (1/RR).

3f. It would still be a saturated model. If we looked at the distribution of LDL across vital status groups, then our model would be finding the probability of high LDL for subjects who survived/died within 5 years, which is different from what we currently have for b and c. The rate ratio would be different as well.

Problem 4 = 14/35

4a. **Method**: Linear regression was used to look at the association between 5-year mortality and LDL levels in a sample of 735 elderly subjects between the ages of 65 and 99. 5-year survival was dichotomized and LDL level was continuous. A significance level of 0.05 was used.

**Results**: We estimate the risk difference of dying within 5 years for two groups differing in LDL by 1 mg/dL to be -0.10%, which is statistically significant (p=0.012). From the 95% confidence interval, it would not be unusual if the risk difference of dying within 5 years is between -0.18% and -0.02%.

4b. **Method**: Poisson regression was used to look at the association between 5-year mortality and LDL levels in a sample of 735 elderly subjects between the ages of 65 and 99. 5-year survival was dichotomized and LDL level was continuous. A significance level of 0.05 was used.

**Results**: We estimate the risk ratio of dying within 5 years for two groups differing in LDL by 1 mg/dL to be 0.994, which is statistically significant (p=0.021). From the 95% confidence interval, it would not be unusual if the risk ratio of dying within 5 years were between 0.988 and 0.999.

4c. **Method**: Logistic regression was used to look at the association between 5-year mortality and LDL levels in a sample of 735 elderly subjects between the ages of 65 and 99. 5-year survival was dichotomized and LDL level was continuous. A significance level of 0.05 was used.

**Results**: We estimate the odds ratio of dying within 5 years for two groups differing in LDL by 1 mg/dL to be 0.992, which is statistically significant (p=0.012). From the 95% confidence interval, it would not be unusual if the risk ratio of dying within 5 years is between 0.986 and 0.998.

4d. In problems 1-3 of this homework, we got an insignificant p-value when dichotomizing LDL. When we had a continuous LDL variable in part 4, we got a significant p-value. Problems 2 and 4 in HW2 also gave us significant p-values when we kept our LDL continuous. Keeping LDL continuous would be better *a priori*, since dichotomizing leads to loss of information as it is only comparing two groups.