BIOST 518

Homework #3

Due Date: January 27, 2014

Question 1:

Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)

(a) Is this a saturated regression model? Explain your answer.

**ANSWER: This is a saturated model. There are two parameters () and two distinct groups (low and high LDL groups).**

(b) For subjects with low LDL, what is the estimated odds of dying within 5 years? What is the estimated probability of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**ANSWER:**

**Log odds model expressed as:**

**Low LDL is defined as LDL < 160 mg/dL.**

**Using the logit command from STATA:**

**To estimate the odds:**

**EXP(-1.58632 - 0.30723(0)) = 0.205 (low LDL grouped as 0). The estimated odds of dying within 5 years is 0.205 among patients with low LDL.**

**The estimated probability of dying within 5 years among patients with low LDL is (0.205/1+0.205) = 0.17 or 17%.**

**The observed proportion of subjects dying within 5 years with low LDL is 0.1699 (105/(105+513) or 17%. This is exactly the same as that calculated using the logistic regression model because the model is saturated.**

(c) For subjects with high LDL, what is the estimated odds of dying within 5 years? What is the estimated probability of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**ANSWER:**

**Log odds model expressed as:**

**High LDL is defined as LDL >/= 160 mg/dL.**

**Using the logit command from STATA:**

**To estimate the odds:**

**EXP(-1.58632 - 0.30723(1)) = 0.151 (high LDL grouped as 1). The estimated odds of dying within 5 years is 0.151 among patients with high LDL.**

**The estimated probability of dying within 5 years among patients with high LDL is (0.151/1+0.151) = 0.131 or 13.1%.**

**The observed proportion of subjects dying within 5 years with high LDL is 0.1308 (14/(14+93) or 13.1%. This is exactly the same as that calculated using the logistic regression model because the model is saturated.**

(d) Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**ANSWER:**

**Methods:**

**We performed a logistic regression evaluating the association between LDL groups (low LDL and high LDL) and death within 5 years. Low LDL was a binary variable defined as serum LDL < 160 mg/dL. High LDL was defined as serum LDL >/=160 mg/dL. The response variable death within 5 years is a binary variable. We sought to compare the odds of response across all LDL groups using an odds ratio. As association will exist if the slope () is nonzero; the odds of death within 5 years will be different across different LDL groups. Regression parameters were estimated using maximum likelihood estimation based on the Binomial distribution for death within 5 years. Huber-White sandwich estimate was used to estimate robust standard errors.**

**Results:**

**Of the 735 subjects in the study, 725 had data for LDL levels. 119 (16.4%) subjects died within 5 years and 606 (83.6%) did not die within 5 years. From logistic regression analysis, we estimate that for patients with high LDL the odds of death within 5 years is 36% lower compared to patients with high LDL. However, this estimate is not statistically significant (P=0.316, two-sided).**

**A 95% confidence interval suggests that this observation is not unusual if patients with low LDL have odds of death within 5 years that was anywhere from 60% lower or 34% higher than patient with high LDL.**

**Compared to Problem 5 in HW#1, the P-values were similar. The P-value for Problem 5 was 0.314 (chi squared test, two-sided), which is 0.6% lower than what was observed in the logistic regression model. The chi-square test is derived from the score test of the logistic regression; therefore, we would expect to get exact results with both methods.**

**In Homework 1, Problem #6, the odds ratio for the comparison of the high LDL group to the low LDL group was 0.735 (95% CI: 0.404, 1.34) using Wald (Woolf) test (P-value is 0.314, two-sided). In the logistic regression model, the odds ratio for the comparison of the high LDL group to the low LDL group was 0.735 (95% CI: 0.404, 1.34) (P-value 0.316, two-sided). The 95% CI are derived from maximum likelihood estimator methods which is based on the Wald test. Since the model is saturated, we would expect to get the exact same result.**

(e) How would the answers to parts a-c change if I had instead asked you to fit a logistic regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**ANSWER:**

**Log odds regression model expressed as:**

**Part (a): In part a, the model is still saturated because there are two parameters () and two distinct groups (low and high LDL groups).**

**Part (b): In part b, all the estimates are exactly the same. The model intercept and slope are different as expressed by the log odds model.**

**Using the logit command from STATA:**

**To estimate the odds:**

**EXP(-1.8935 + 0.30723(1)) = 0.205 (low LDL grouped as 1). The estimated odds of dying within 5 years is 0.205 among patients with low LDL.**

**The estimated probability of dying within 5 years among patients with low LDL is (0.205/1+0.205) = 0.17 or 17%.**

**The proportion of subjects dying within 5 years with low LDL is 0.1699 (105/(105+513) or 17%. This is exactly the same as that calculated using the logistic regression model because the model is saturated.**

**Part (c): In part c, all the estimates are exactly the same. The model intercept and slope are different as expressed by the log odds model.**

**Using the logit command from STATA:**

**To estimate the odds:**

**EXP(-1.8935 + 0.30723(0)) = 0.151 (high LDL grouped as 0). The estimated odds of dying within 5 years is 0.151 among patients with high LDL.**

**The estimated probability of dying within 5 years among patients with high LDL is (0.151/1+0.151) = 0.131 or 13.1%.**

**The proportion of subjects dying within 5 years with high LDL is 0.1308 (14/(14+93) or 13.1%.**

**Different indicator of survival:**

**If we had used a different indicator of survival, the odds would be the inverse of what we observed. The log odds model would be expressed as:**

**Part (a): The model is still a saturated model because there are two parameters () and two distinct groups (high versus low LDL)**

**Part (b):**

**EXP(1.8935 - 0.30723(1)) = 4.89 (low LDL grouped as 1). The estimated odds of surviving at least 5 years is 4.89 among patients with high LDL. This is the inverse of the answer from Part (b) (1/0.151).**

**The estimated probability of surviving at least 5 years among patients with low LDL is (4.89/1+0.4.89) = 0.83 or 83%.**

**The proportion of subjects surviving at least 5 years with low LDL is 0.83 (513/(513+105) or 83%.**

**Part (c):**

**To estimate the odds:**

**EXP(1.8935 - 0.30723(0)) = 6.64 (high LDL grouped as 0). The estimated odds of surviving at least 5 years is 6.64 among patients with high LDL. This is the inverse of the answer from Part (b) (1/0.151).**

**The estimated probability of surviving at least 5 years among patients with low LDL is (6.64/1+0.6.64) = 0.869 or 87%.**

**The proportion of subjects surviving at least 5 years with low LDL is 0.869 (93/(93+14) or 87%.**

(f) In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a logistic regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**ANSWER:**

**Consider a model where the predictor is Death within 5 years and the response is high LDL.**

**,**

**where death within 5 years = 1 and did not die within 5 years = 0.**

**Part (a): This is still a saturated mode because we still have parameters () and two distinct groups (Died and did not die within 5 years).**

**Part (b):**

**The estimated odds of having high LDL if the subject did not die within 5 years:**

**EXP(-1.7077 - 0.3072(0)) = 0.181. The odds of having low LDL is 0.181 among those who did not die within 5 years.**

**The estimated probability of having high LDL among those who did not die within 5 years is 0.153 (0.181/(1+0.181)) or 15.3%.**

**The observed proportion of patients with high LDL among those who did not die within 5 years is 15.3% (14/(14+105)).**

**Part (c):**

**The estimated odds of having high LDL if the subject died within 5 years:**

**EXP(-1.7077 - 0.3072(1)) = 0.133. The odds of having high LDL is 0.133 among those who died within 5 years.**

**The estimated probability of having high LDL among those who died within 5 years is 0.118 (0.507/(1+0.507)) or 11.8%.**

**The observed proportion of patients with high LDL among those who died within 5 years is 11.8% (14/(14+105)).**

Question 2:

Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the differences in the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)

(a) Is this a saturated regression model? Explain your answer.

**ANSWER:**

**This is a saturated model. There are two parameters () and two distinct groups (low and high LDL groups).**

**We are looking at the difference in probability of death differing across LDL groups as our summary measure (or the Risk Difference). Since we are looking at the average difference in probability, we will need to use a linear regression. We use the linear regression model expressed as:**

**High LDL = 1 and Low LDL = 0.**

(b) For subjects with low LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**ANSWER:**

**The odd of dying within 5 years among those with low LDL is 0.205. I estimated this using the linear regression model:**

**0.1699029 - 0.0390618(0) = 0.1699 (low LDL grouped as 0). Odds is calculated using the formula: P/(1-P) or 0.1699/(1-0.1699) = 0.205.**

**The estimated probability of subjects dying within 5 years among those with low LDL is 17%.**

**The observed proportion of subjects dying within 5 years with low LDL is 0.1699 (105/(105+513) or 17%. This is exactly the same as that calculated using the linear regression model because the model is saturated.**

(c) For subjects with high LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**ANSWER:**

**The estimated odds of dying within 5 years among those with high LDL is 0.1505. I estimated this using the linear regression model:**

**0.1699029 - 0.0390618(1) = 0.1308 (low LDL grouped as 0). Odds is calculated using the formula: P/(1-P) or 0.1308/(1-0.1308) = 0.1505.**

**The estimated probability of subjects dying within 5 years among those with low LDL is 13.1%.**

**The observed proportion of subjects dying within 5 years with high LDL is 0.1308 (14/(14+93) or 13.1%. This is exactly the same as that calculated using the logistic regression model because the model is saturated.**

(d) Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**ANSWER:**

**Methods:**

**We performed a linear regression evaluating the association between differences in probability of death within 5 years across LDL groups (low LDL and high LDL). We treat LDL groups as a continuous variable, even though it is a binary variable. (Low LDL was a binary variable defined as serum LDL < 160 mg/dL. High LDL was defined as serum LDL >/=160 mg/dL).**

**The response variable death within 5 years is a binary variable. We sought to compare the difference in probability of response across all LDL groups. As association will exist if the slope () is nonzero; the probability of death within 5 years will be different across different LDL groups. We performed the linear regression with the presumption of equal variances across groups. Regression parameters were estimated using ordinary least squares estimation (assuming homoscedasticity) corresponding to maximum likelihood estimation.**

**Results:**

**Of the 735 subjects in the study, 725 had data for LDL levels. 119 (16.4%) subjects died within 5 years and 606 (83.6%) did not die within 5 years. From the linear regression model, the probability of death within 5 years among patients in the high LDL group was 3.9% lower relative to patients in the low LDL group (P-value = 0.315, two-tailed). This is also the slope of the model (). A 95% confidence interval suggests that this observation is not unusual if the true difference in probability of death within 5 years among patients with high LDL and low LDL were between 11.5% lower and 3.7% higher. Because the P-value is 0.315, we do not have enough evidence to reject the null hypothesis that there is no linear trend in the probability of death within 5 years across LDL groups.**

**Compared to Problem 5 in HW#1, the difference in probability of death within 5 years across different LDL groups is the same (3.91%) to what was estimated using the linear regression. The P-values are the same. The P-value for Problem 5 was 0.314 (chi squared test, two-sided). However, the confidence interval is different. For Problem #6, the 95% confidence interval is from 10.9% absolute probability of survival to a 3.14% absolute probability of survival in the high LDL group compared to the low LDL group. The 95% CI for the linear regression model is wider compared to Problem #6. The CI for the linear regression model is based on the pooled standard deviation or RMSE.**

**In Homework 1, Problem #6, the odds ratio for the comparison of the high LDL group to the low LDL group was 0.735 (95% CI: 0.404, 1.34) using Wald (Woolf) test (P-value is 0.314, two-sided). In the linear regression model, the odds ratio for the comparison of the high LDL group to the low LDL group was 0.737 (0.151/0.205). This is very similar to the OR calculated in Problem #6, HW1.**

(e) How would the answers to parts a-c change if I had instead asked you to fit a regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**ANSWER:**

**We use the linear regression model expressed as:**

**Low LDL = 1 and High LDL = 0.**

**Part (a): This is a saturated model. There are two parameters () and two distinct groups (low and high LDL groups).**

**Part (b):**

**0.1308411 + 0.0390618(1) = 0.1699. (Low LDL group = 1)**

**The estimated odds of death within 5 years among patients who have low LDL is 0.205 (0.1699/(1-0.1699)). This is the same answer as in part (b).**

**The estimated probability of death within 5 years among patients who have low LDL is 17%. The is the same answer as in part (b).**

**The observed probability of death within 5 years among patients who have low LDL is 17% or (105/(105+513)).**

**Part (c):**

**The estimated odds of dying within 5 years among those with high LDL is 0.1505. I estimated this using the linear regression model:**

**0.1308411 + 0.0390618(0) = 0.1308 (high LDL grouped as 0). Odds is calculated using the formula: P/(1-P) or 0.1308/(1-0.1308) = 0.1505. This is the same answer as in part (c).**

**The estimated probability of subjects dying within 5 years among those with low LDL is 13.1%. This is the same answer as in part (c).**

**The observed proportion of subjects dying within 5 years with high LDL is 0.1308 (14/(14+93) or 13.1%. This is the same answer as in part (c).**

**Using an indicator of survival for at least 5 years as the response variable:**

**We use the linear regression model expressed as:**

**Low LDL = 1 and High LDL = 0.**

**Part (a): This is a saturated model. There are two parameters () and two distinct groups (low and high LDL groups).**

**Part (b):**

**0.8691589 - 0.0390618(1) = 0.83. (Low LDL group = 1)**

**The estimated odds of surviving at least 5 years among patients who have low LDL is 4.89 (0.83/(1-0.83)).**

**The estimated probability of survival at least 5 years among patients who have low LDL is 83%.**

**The observed probability of surviving at least 5 years among patients who have low LDL is 83% or (513/(105+513))**

**Part (c):**

**The estimated odds of surviving at least 5 years among those with high LDL is 0.869. I estimated this using the linear regression model: 0.8691589 - 0.0390618(1) = 0.869 (high LDL grouped as 0). Odds is calculated using the formula: P/(1-P) or 0.869/(1-0.869) = 6.63.**

**The estimated probability of subjects dying within 5 years among those with low LDL is 86.9%. This is the same answer as in part (c).**

**The observed proportion of subjects dying within 5 years with high LDL is 0.869 (93/(14+93) or 86.9%.**

(f) In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**ANSWER:**

**Consider a model where the predictor is Death within 5 years and the response is high LDL.**

**where death within 5 years = 1 and did not die within 5 years = 0.**

**Part (a): This is still a saturated mode because we still have two parameters () and two distinct groups (Died and did not die within 5 years).**

**Part (b):**

**0.1534653 - 0.0358183(0) = 0.153 or 15.3% (Did not die within 5 years is grouped as 0).**

**The estimated odds of having high LDL if the subject did not die within 5 years:**

**0.181 (0.153/(1-0.153)). The odds of having high LDL is 0.181 among those who did not die within 5 years.**

**The estimated probability of having high LDL among those who did not die within 5 years is 15.3%.**

**The observed proportion of patients with high LDL among those who did not die within 5 years is 15.3% (14/(14+105)).**

**Part (c):**

**0.1534653 - 0.0358183(1) = 0.118. (Did not die within 5 years is grouped as 1).**

**The estimated odds of having high LDL if the subject died within 5 years:**

**The odds of having high LDL is 0.133 (0.118/(1-0.118) among those who died within 5 years.**

**The estimated probability of having high LDL among those who died within 5 years is 11.8%.**

**The observed proportion of patients with high LDL among those who died within 5 years is 11.8% (14/(14+105)).**

Question 3:

Perform a statistical regression analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the ratios of the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL > 160 mg/dL). In your regression model, use an indicator of death within 5 years as your response variable, and use an indicator of high LDL as your predictor. (Only give a formal report of the inference where asked to.)

(a) Is this a saturated regression model? Explain your answer.

**ANSWER:**

**This is still saturated mode because we have two parameters () and two distinct groups (Died and did not die within 5 years).**

**We are looking at the ratios of the probability of death within 5 years as the summary measure. Therefore, we will need to construct a Poisson regression model. This is akin to a risk ratio. We use the Poisson regression model expressed as:**

**where High LDL is grouped as 1 and Low LDL is grouped as 0.**

(b) For subjects with low LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**ANSWER:**

**log rate = -1.7725 – 0.2612(High LDL = 0)**

**EXP(-1.7725) = 0.1699**

**The estimated probability of dying within 5 years among patients with low LDL is 17%.**

**The estimated odds of dying within 5 years among patients with low LDL is 0.205 (0.1699/(1-0.1699)).**

**The observed proportion of dying within 5 years among patients with low LDL is 17% (105/(105+5133)). This is exactly the same as that calculated using the Poisson model because the model is saturated.**

(c) For subjects with high LDL, what is the estimated probability of dying within 5 years? What is the estimated odds of dying within 5 years? How do these estimates compare to the observed proportion of subjects with low LDL dying within 5 years?

**ANSWER:**

**log rate = -1.7725 – 0.2612(High LDL = 1)**

**EXP(-2.0337) = 0.1309**

**The estimated probability of dying within 5 years among patients with high LDL is 13.1%.**

**The estimated odds of dying within 5 years among patients with high LDL is 0.151 (0.1309/(1-0.1309)).**

**The observed probability of dying within 5 years among patients with high LDL is 13.1% (14/(14+93)). This is exactly the same as that calculated using the Poisson model because the model is saturated.**

(d) Give full inference regarding the association between 5 year mortality and high LDL levels. How does this differ from the inference that was made on problems 5 and 6 of homework #1? What is the source of any differences?

**ANSWER:**

**Methods:**

**We performed a Poisson regression evaluating the association between ratios of the probability of death within 5 years across LDL groups (low LDL and high LDL). (Low LDL was a binary variable defined as serum LDL < 160 mg/dL. High LDL was defined as serum LDL >/=160 mg/dL).**

**The response variable death within 5 years is a binary variable. We sought to compare the ratio in probability of response between high and low LDL groups. As association will exist if the slope () is nonzero; the probability of death within 5 years will be different across different LDL groups. Regression parameters were estimated using maximum likelihood estimation based on the Poisson distribution of death within 5 years. Variability of the regression parameter estimates was based on the mean-variance relationship.**

**We also assume that the response variable is non-negative, which it is.**

**Results:**

**Of the 735 subjects in the study, 725 had data for LDL levels. 119 (16.4%) subjects died within 5 years and 606 (83.6%) did not die within 5 years. From the Poisson regression analysis, we estimated that the ratios of the probability of death within 5 years across among patients in the high LDL group relative to patients in the low LDL group is 0.77 (EXP(-0.2612)) with a two-tailed P-value of 0.359. Since the P-value is 0.359, we do not have enough evidence to reject the null hypothesis that there is no association between dying within 5 years and LDL groups. Based on the 95 % CI, this observation is not unusual if the ratios of probability among patients in the High LDL group relative to the Low LDL group were anywhere from 0.44 and 1.35.**

**In HW#1 Problem 5, the probabilities of dying within 5 years among high LDL and low LDL groups were 13.1% and 17%, respectively. These are the same results as the above derivation from the Poisson regression. The two-sided P-value was 0.359, which is larger than the one obtained from HW1 (two-sided P-value=0.314).**

**In HW#1 Problem 6, the odds of dying within 5 years among high LDL and low LDL groups were 0.151 and 0.204, respectively. These are the same results are the above derivation from the Poisson regression.**

**Differences in P-value were observed probably due to the mean-variance relationship in the Poisson regression analysis. Poisson regression analysis uses maximum likelihood estimation to find parameter estimates. The 95% CI is derived from the Wald chi-square statistic. Poisson undergoes an iterative search for the best fitting model.**

(e) How would the answers to parts a-c change if I had instead asked you to fit a regression model using the indicator of death within 5 years as your response variable, but using an indicator of low LDL as your predictor? What if we had used an indicator of survival for at least 5 years as the response variable?

**ANSWER:**

**We are looking at the ratios of the probability of death within 5 years as the summary measure. Therefore, we will need to construct a Poisson regression model. This is akin to a risk ratio. We use the Poisson regression model expressed as:**

**where High LDL is grouped as 0 and Low LDL is grouped as 1.**

**Part (a): This is still saturated mode because we have two parameters () and two distinct groups (Died and did not die within 5 years).**

**Part (b):**

**log rate = -2.034 + 0.2612(Low LDL = 1)**

**EXP(-1.7728) = 0.1699**

**The estimated probability of dying within 5 years among patients with low LDL is 17%.**

**The estimated odds of dying within 5 years among patients with low LDL is 0.205 (0.1699/(1-0.1699)).**

**The observed proportion of dying within 5 years among patients with low LDL is 17% (105/(105+5133)). This is exactly the same as that calculated using the Poisson model because the model is saturated.**

**Part (c):**

**log rate = -2.034 + 0.2612(Low LDL = 0)**

**EXP(-2.034) = 0.1308**

**The estimated probability of dying within 5 years among patients with high LDL is 13.1%.**

**The estimated odds of dying within 5 years among patients with high LDL is 0.151 (0.1309/(1-0.1309)).**

**The observed probability of dying within 5 years among patients with high LDL is 13.1% (14/(14+93)). This is exactly the same as that calculated using the Poisson model because the model is saturated.**

**Using an indicator of survival for at least 5 years as the response variable:**

**We use the Poisson regression model expressed as:**

**Part (a): This is still saturated mode because we have two parameters () and two distinct groups (Died and did not die within 5 years).**

**Part (b):**

**log rate = -0.1402 – 0.04598(Low LDL = 1)**

**EXP(-0.1862) = 0.83**

**The estimated probability of surviving at least 5 years among patients with low LDL is 83%.**

**The estimated odds of surviving at least 5 years among patients with low LDL is 4.89 (0.83/(1-0.83)).**

**The observed proportion of surviving at least 5 years among patients with low LDL is 83% (513/(105+513)).**

**Part (c):**

**log rate = -0.1402 – 0.04598(Low LDL = 0)**

**EXP(-0.1492) = 0.8691**

**The estimated probability of dying within 5 years among patients with high LDL is 86.9%.**

**The estimated odds of dying within 5 years among patients with high LDL is 6.64 (0.8691/(1-0.8691)).**

**The observed probability of dying within 5 years among patients with high LDL is 86.9% (93/(14+93)).**

(f) In parts a-d of this problem, we described the distribution of death within 5 years across groups defined by LDL level. What if we fit a regression model mimicking the approach used in problems 1 – 4 of homework #2, where we described the distribution of LDL across groups defined by vital status? How would our answers to parts a-c change?

**ANSWER:**

**We are looking at the ratios of the probability of High LDL as the summary measure. Therefore, we will need to construct a Poisson regression model. This is akin to a risk ratio. We use the Poisson regression model expressed as:**

**where Death within 5 years is grouped as 1 and Not dead within 5 years is grouped as 0.**

**(a)**

**This is still saturated mode because we have two parameters () and two distinct groups (Died and did not die within 5 years).**

**(b)**

**log rate = -1.8743 – 0.2658(Dead within 5 years = 0)**

**EXP(-1.8743) = 0.1534**

**The estimated probability of having High LDL among patients not dead within 5 years is 15.3%.**

**The estimated odds of having High LDL among patients with not dead within 5 years is 0.181 (0.1534/(1-0.1534)).**

**The observed proportion of having High LDL among patients Dead within 5 years is 15.3% (93/(93+513)).**

**(c)**

**log rate = -1.8743 – 0.2658(Dead within 5 years = 1)**

**EXP(-2.1401) = 0.1176**

**The estimated probability of having High LDL among patients Dead within 5 years is 11.8%.**

**The estimated odds of having High LDL among patients Dead within 5 years is 0.133 (0.1176/(1-0.1176)).**

**The observed proportion of High LDL among patients Dead within 5 years is 11.8% (14/(14+105)).**

Question 4:

Perform a regression analysis of the distribution of death within 5 years across groups defined by the continuous measure of LDL. (In all cases we want formal inference.)

(a) Evaluating the association between 5-year mortality and LDL using risk difference (RD: difference in probabilities)

**ANSWER:**

**Methods:**

**We performed a linear regression evaluating the association between differences in probability of death within 5 years across LDL levels. We treat LDL groups as a continuous variable. The response variable death within 5 years is a binary variable. We sought to compare the difference in probability of response across LDL levels. An association will exist if the slope () is nonzero; where the probability of death within 5 years will be different across different LDL levels. We performed the linear regression with the presumption of equal variances across groups. Parameter estimates were calculated using ordinary least square method corresponding to maximum likelihood estimation. Variance was presumed to be equal and homoscedastic.**

**Linear regression expressed as:**

**Results:**

**Of the 735 subjects in the study, 725 had data for LDL levels. 119 (16.4%) subjects died within 5 years and 606 (83.6%) did not die within 5 years. From the linear regression, we estimate that for each difference in LDL level by 1 mg/dL, the difference in probability of death within 5 years is 0.1% lower in the higher LDL group; and this estimate is statistically significant (two-tailed P=0.012). A 95% CI suggests that this observation is not unusual if a group that is 1 mg/dL higher might have a difference in risk of dying within 5 years that is between 0.02% lower and 0.18% lower.**

(b): Evaluating the association between 5-year mortality and LDL using risk ratio (RR: ratios of probabilities)

**ANSWER:**

**Methods:**

**We performed a Poisson regression evaluating the association between ratios of the probability of death within 5 years across LDL levels. The response variable death within 5 years is a binary variable. We sought to compare the ratios in probability of response between LDL levels. As association will exist if the slope () is nonzero; the probability of death within 5 years will be different across different LDL levels. Parameter estimates were calculated using maximum likelihood estimation based on the Poisson distribution for death within 5 years. Variability of the regression parameter estimates are based on the mean-variance relationship dictated by the Poisson distribution.**

**Poisson regression expressed as:**

**Results:**

**Of the 735 subjects in the study, 725 had data for LDL levels. 119 (16.4%) subjects died within 5 years and 606 (83.6%) did not die within 5 years. From the Poisson regression, we estimate that for each 1 mg/dL difference in LDL, the risk ratio of death within 5 years was 0.64% lower in the higher LDL group. This estimate is statistically significant; the two-sided P-value was 0.021. A 95% CI suggests that this observation is not unusual if a group that is 1 mg/dL higher might have a risk ratio for dying within 5 years was between 0.1% lower and 1.2% lower.**

Part (c): Evaluating the association between 5-year mortality and LDL using odds ratio (OR: ratios of odds)

**ANSWER:**

**Methods:**

**We performed a logistic regression evaluating the association between odds of death within 5 years across LDL levels. The response variable death within 5 years is a binary variable. We sought to compare the odds of response between LDL levels. As association will exist if the slope () is nonzero; the odds of death within 5 years will be different across different LDL levels. Regression parameters were estimated using maximum likelihood estimation based on a Binomial distribution for death within 5 years. 95% CI were derived from the Wald-based chi-square test. Huber-White sandwich estimator was used to calculate the standard errors.**

**The log odd model is expressed as:**

**Results:**

**Of the 735 subjects in the study, 725 had data for LDL levels. 119 (16.4%) subjects died within 5 years and 606 (83.6%) did not die within 5 years. From logistic regression analysis, we estimated that for each 1 mg/dL difference in LDL, the odds of dying within 5 years is 0.77% lower in the higher LDL group, and this estimate is statistical significant (two-tailed P=0.019). A 95% CI suggests that this observation is not unusual if a group that is 1 mg/dL higher might have odds of dying within 5 years that was anywhere from 0.13% lower and 1.42% lower than the lower LDL groups.**

(d) How do your conclusions about such an association from this model compare to your conclusions reached in problems 1-3 of this homework and problems 2 and 4 of homework #2? Which analyses would you prefer a priori.?

**ANSWER:**

**Conclusions were similar to Problems 1-3 of this homework. The general association is that having higher LDL implies lower risk of dying within 5 years. This was observed with the probability, odds, and risk ratio.**

**Problems 2 and 4 from HW#2 evaluated the distribution of LDL across vital status. Therefore, the intercept provided the estimated mean serum LDL in a population who would not die within 5 years. It does not provide us with the probability of death within 5 years.**

**I would prefer to use the Poisson regression analysis to get the ratios of proportions since this is a cohort study and not a case-control study. I wouldn’t be interested in OR unless the event of death is rare. In this case, we may still use the logistic regression, but the Poisson gives us RR, which is preferable.**