**Biostats 518 Homework #3**

1. Question 1
   1. We have two groups (those who died within 5 years and those who didn’t). Since we have 2 parameters in our model (intercept and slope), this is a saturated model.
      1. E(Group|LDL > 160) = .169 - .039X
   2. The odds of dying within 5 years is .16/.13 = 1.23. The estimated probability of dying within 5 years is 0.169. The proportion of subjects with low ldl who died within 5 years is 0.11/0.15, or 73%. The odds seem to indicate a higher risk and the probability seems to indicate a lower risk.
   3. The odds of dying within 5 years for the high LDL group is .13/.16 = 0.8. The estimated probability of dying within 5 years is 0.13. The proportion of subjects with low ldl who died within 5 years is 0.11/0.15, or 73%. The odds seem to indicate a lower risk and the probability seems to indicate a lower risk as well.
   4. Full inference:
      1. **Method:** Generating a linear regression with the group (died within 5 years or not) as the predictor and whether or not the group member had “high” LDL as the response variable. Assuming homostadicity.
      2. **Inference:** 5-year all cause mortality does not seem to be strongly associated with high LDL. Based on the model, having lower LDL reduces the chance of being in the deadin5 group. However, having a higher LDL only decreases the chances by 0.04. Hardly significant. Questions 5 and 6 on Homework #1 came to the same conclusions. However, they used Chi-square test and Fisher’s Exact test to get that answer.
   5. A logistic regression looks like this: E(Group|LDL>160) = -1.586 – 0.307X. My answers to a-c would be the same because those characteristics are independent of the regression model. If we had used an indicator of survival for at least 5 years as the response variable, the answers would the inverse of what we have now, but the conclusions would still be the same.
   6. Again, the answers wouldn’t change because we are using a breakdown of the data that still represents the same thing. Even though we have categories numbers 1, 2, and 3, they still represent the LDL intervals they have always represented—the data is intact. Our answers to parts a-c would not change unless the definitions of “high” and “low” were changed.
2. Question 2
   1. We have two defining groups, high LDL and not high LDL. Since we have two parameters in our model, this model is saturated.
      1. E(Group|LDL>160) = -1.586 – 0.307X
   2. The odds of dying within 5 years is -1.89/-1.56 = 1.19. The estimated probability of dying within 5 years is 0.169. The proportion of subjects with low ldl who died within 5 years is 0.11/0.15, or 73%. The odds seem to indicate a higher risk and the probability seems to indicate a lower risk.
   3. The odds of dying within 5 years for the high LDL group is -1.56/-.189 = 0.82. The estimated probability of dying within 5 years is 0.13. The proportion of subjects with low ldl who died within 5 years is 0.11/0.15, or 73%. The odds seem to indicate a lower risk and the probability seems to indicate a lower risk as well.
   4. Full inference:
      1. **Method:** Generating a logistic regression with the group (died within 5 years or not) as the predictor and whether or not the group member had “high” LDL as the response variable. Assuming homostadicity.
      2. **Inference:** 5-year all cause mortality does not seem to be strongly associated with high LDL. Based on the model, having lower LDL reduces the chance of being in the deadin5 group. Questions 5 and 6 on Homework #1 came to the same conclusions. However, they used Chi-square test and Fisher’s Exact test to get that answer.
   5. My answers to a-c would be the same because those characteristics are independent of the model. If we had used an indicator of survival for at least 5 years as the response variable, the answers would the inverse of what we have now, but the conclusions would still be the same.
   6. For the most part, the answers wouldn’t change because we are using a breakdown of the data that still represents the same thing. There may be slight differences in the answer output by the model, but there will likely not be enough of a difference to come to a different conclusion.
3. Question 3
   1. We have two groups of interest (dead within 5 years and survived) and two parameters (slope and intercept), so we have a saturated model.
      1. E(Group|LDL>160) = -1.77 - .26X
   2. For those with low LDL, the probability of dying within 5 years is -1.77/-2.03 = 87%. The proportion of subjects with low ldl who died within 5 years is 0.11/0.15, or 73%. The probability of dying seems to be higher this time.
   3. For those with high LDL, the probability of dying within 5 years is -2.03/-1.77 = 14% higher than the low LDL group. The proportion of subjects with low ldl who died within 5 years is 0.11/0.15, or 73%. This seems to be a higher proportion of high-risk, low-ldl subjects than what we saw earlier.
   4. Full inference:
      1. **Method:** Generating a poisson regression with the cholesterol group (“high” or “not high””) as the predictor and whether or not the subject died within 5 years or survived more than 5 years as the response variable. Assuming homostadicity.
      2. **Inference:** 5-year all cause mortality does seem to be strongly associated with high LDL There is a 14% higher absolute risk for those with high LDL. On questions 5 and 6 on homework #1, we basically ran a test of independence and the results were the opposite of our conclusion here.
   5. The answers for a-c would me more conservative because we assume a normal distribution with a simple regression. Poisson regression is more robust and takes into account the fact that the distribution might not be normal. If we used survival as the response variable, we would have the same conclusion (but with inverse analysis results).
   6. Since the null hypothesis was rejected under this model, we may see a stronger trend for the higher LDL group members and weaker trends for the other two categories of LDL groups.
4. Question 4
   * 1. **Method:** Logistic regression with LDL as the predictor and deadin5 group as the response variable.
        1. E(Group|LDL) = 0.294 -.001X
        2. .134 = 0.294 -.001\*(160)
        3. .174 = 0.294 -.001\*(120)
     2. **Inference:** There seems to be a little difference in relationship between groups in this model. The model is saying that those with higher LDL values have a lower risk of being in the deadin5 group. There seems to be an absolute risk difference of 40% in the scenario I painted above.
     3. **Method:** Poisson regression with LDL as the predictor and deadin5 group as the response variable.
        1. E(Group|LDL) = -1.01 -.006X
        2. -1.97 = -1.01 -.006 \*(160)
        3. -1.73 = -1.01 -.006\*(120)
     4. **Inference:** The ratio of high LDL to not high LDL is -1.97/-1.73 = 1.13. This is saying that those with high LDL values are at a 13% higher probabilistic risk of being in the deadin5 group.
     5. **Method:** Logistic regression with LDL as the predictor and deadin5 group as the response variable.
        1. E(Group|LDL) = -.67 -.007X
        2. -1.79 = -.67 -.007\*(160)
        3. -1.51 = -.67 -.007\*(120)
     6. **Inference:** There seems to be a little difference in relationship between groups in this model. The model is saying that those with higher LDL values have a lower risk of being in the deadin5 group.