**Biostats 518 Homework #3 43/108**

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. 3/3
      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. 1/3 The estimated probability of death given low ldl is correct. I believe the rest of the analysis was trying to predict LDL status based on death however it was intended to be the otherway around. Also, note that odds is simply (# in group dead)/(#number in group alive). It appears you were trying to compute an odds ratio. This problem was also looking to utilize logistic regression for the estimates.
   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. 1/3, the comments from part b also relate to this problem.
   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.

We want to utilize a logistic regression model here. Also, state whether your model is using a two or one sided pa vlue. Also, the response variable should be death at 5 years and the predictor high LDL. 0/5

* + 1. **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. 2/5, more explanation of the results o the analysis (p value, confidence intervals) is necessary for a statistical inference as well as the comparison. Partial credit was given for the interpretation despite the switching of the variables.
  1. 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. 2/3 Expand more on the actual models rather than just the end conclusion.
  2. 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. 1/3 The answer indeed would not change, however the reason relates to the calculation of an odds ratio, which somewhat uniquely has this property. The slope of the two models would be equivalent but not the intercept

1. 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. 3/3
      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.

The negative numbers should not be present, perhaps you forgot to exponentiate your values? The estimated probability of dying is correct however the proportion of subjects with low LDL who died is off. It may be that you were calculating the probability of low LDL given death. 1/3

* 1. 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.

Similar comments to part b above. 1/3

* 1. 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.

The response and predictor variables got mixed up for this analysis. Also be sure to specify if you used a two sided test. 1/5

* + 1. **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.
    2. **See the response for question 1.** Make sure to include your point estimates as well. 1/5
  1. 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. 2/3 Make sure to include the relationshipsbetween the models.
  2. 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. If there are slight differences be sure to mention what those differences are. These models are quite different as we are predicting different responses and the consistency does not hold true as it does for odds ratios. The statistical association between the variables should remain the same as you alluded to correctly. 1/3

1. 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. 3/3
      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.

I believe the exponentiation step was left out here as well. Also your formula was for the odds ratio between the two groups, not the odds for one of the groups. 1/3

* 1. 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. Similar to part b 1/3
  2. 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.

See comments in other methods sections 3/5.

* + 1. **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.

**Be sure to include more values such as p values and confidence intervals to fully support your statistical inference 2/5.**

* 1. 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).

I believe the question was misunderstand, the key was re-parameterization. The key provides a much more detailed description then I could provide. 0/3

* 1. 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.

The comments looked for on the key were regarding “reversing” the model. Although the estimates would be difference we would not be surprised to find a similar association. 0/3

1. 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)

For risk differences, a linear regression model would be preferred as it models the difference between two risk groups rather than the odds ratio between them, which is the analysis you ran, just the wording is off. Be sure to state what type of tests this utilizes and whether the standard errors are robust or not. 2/5

* + 1. **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.

Be sure to include a p value and confidence interval in this analysis to determine whether the findings are significant. Also, try to avoid using variable names in your description as it may be difficult for someone with little background in our homeworks to understand what you mean. Also, likely a typo, but note that the difference you show is 4%. The analysis seems to be correct, just expand on the findings. 1/5

* + 1. **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)

Make sure to state what type of estimates are to be used for the standard errors and whether the test is two sided or not.

3/5

* + 1. **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.

Make sure to exponentiate your estimates. Also include p values and confidence intervals for your esimates to make a statistical inference. 2/5

* + 1. **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)

See above comments for the methods. 3/5

* + 1. **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.

Similar comments to the inference above. 2/5

Part d 0/3