BST 210 HOMEWORK #6

Due 11:59pm, Tuesday, November 12, 2019

***Please be sure to submit your assignment by 11:55ish pm (or before) to prevent any glitches in the upload from precluding your timely submission.**
***Please work well in advance, getting help during office hours and labs, as there will be no extensions given for this assignment, outside of extreme, extenuating circumstances which must be communicated in advance to the primary instructor.**

There are 2 problems, each with various parts, in this homework assignment. Please double check that you have provided a response for each part of both problems, before you submit.

BST 210 Problem set policies:

- *We encourage you to discuss homework with your fellow students (or with the instructor or the TAs), but you must write your own final answers, in your own words.*
- *Please include the appropriate computer output in your solution if that helps you to answer a question, but be sure to interpret your findings in words – submitting only output is not sufficient for full credit.*
- *Homework assignments will not be accepted late (other than for extreme emergency, but the primary instructor must be reached in advance).*
- *Be complete in your responses; not verbose, to get full scores.*
- *All homework must be submitted online via Canvas by 11:59pm on Tuesday.*

Problem 1

Consider again the Framingham Heart Study dataset. Suppose we are interested in looking at a three-level outcome for *incidence* of *either* coronary heart disease (with the subject still alive) *or* death, compared to a reference group of subjects who neither died nor had coronary heart disease in the follow-up period. Thus, we want to restrict the analysis to *exclude* subjects with prevalent coronary heart disease (prevchd = 1), and create a three-level outcome consisting of:

1 = no death or coronary heart disease in the follow-up period (reference category)

2 = coronary heart disease in the follow-up period, but the subject remained alive

3 = death from any cause in the follow-up period.

Thus, you will need to use prevchd, anychd, and death to create the outcome variable and sample to use. Using this sample, we will explore some multinomial and ordinal logistic regression models, using participant sex and continuous age as predictor variables. It might be easiest to recode sex to be an indicator for female (i.e., = 1 for female, = 0 for male). There should be 4,240 observations if you are using the Framingham dataset, with no one missing outcome, age, or sex.

Fit four multinomial logistic regression models using age alone, sex alone, both age and sex, and finally age, sex, and their interaction, and answer the following questions:

(a) For the model with age alone, calculate and graph the fitted probabilities for each category as a function of age. Briefly interpret your graph. Also, what is the estimated relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 2 to outcome 1? What is the estimated relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 to outcome 1? A little harder: What is the relative risk ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 to outcome 2?

(b) For the model with sex alone, confirm that the fitted probabilities match those of an outcome × sex tabulation exactly. Also confirm that the estimated relative risk ratios for sex from your model match the relative risk ratios from the tabulation. Note that this would only occur with a "saturated model" like when you only have a single dichotomous predictor as here – this will not happen for continuous covariates, say.

(c) Use a LRT to decide between the models including both age and sex and the model including age, sex, and their interaction. What do you conclude? Are there any other models you might recommend fitting next?

Problem 2

Now, fit four ordinal logistic regression models using age alone, sex alone, both age and sex, and finally age, sex, and their interaction, and answer the following questions:

(a) For the model with age alone, what is the estimated odds ratio and 95% CI for the effect of 10 years of age when comparing outcome 3 vs. outcome 1 and 2 (combined)? Also, what is the estimated odds ratio and 95% CI for the effects of 10 years of age when comparing outcome 2 and 3 (combined) vs. outcome 1?

(b) For the model with age alone, is the proportional odds assumption satisfied or rejected? Let's explore this further by these additional looks at the data: First, create a binary outcome variable that equals 1 when you are in category 3 and

equals 0 when you are in category 1 or 2. Run a logistic regression model using age to predict this binary outcome. Second, create a new binary outcome variable that equals 1 when you are in category 2 or 3 and equals 0 when you are in category 1. Again, run a logistic regression model using age to predict this new binary outcome. If the proportional odds assumption holds, we would expect that the two beta coefficients for age in these two models would be close to each another. What happens in this example? (Do the CI's for the age beta coefficients overlap or not?) Given this comparison of the beta coefficients, do you believe the proportional odds model assumption holds or not for the ordinal logistic regression model with age alone?

(c) Now focus on the model with sex alone. Is the proportional odds assumption satisfied or rejected? Let's explore this further by these additional looks at the data: Consider again the outcome × sex tabulation as above. With the ordinal model, we need to tabulate category 1 vs. 2 and 3 (combined) and category 1 and 2 (combined) vs. category 3. If the proportional odds assumption is satisfied, we should feel comfortable with a common odds ratio estimate for these two categorizations. What are the associated odds ratio estimates for sex to predict these categorizations based on hand calculations? How do these compare with the ordinal logistic regression-based odds ratio estimate for sex? Given your comparison of these odds ratio estimates, do you believe the proportional odds model assumption holds or not for the ordinal logistic regression model with sex alone? (One could also perform the pair of logistic regressions as in 2 (b) with sex as the only predictor and compare the beta coefficients for sex in these two models. Try that if you like.) Finally, is this ordinal logistic regression model saturated or not? Defend your answer.

(d) Do we have any evidence that the age × sex interaction is needed for ordinal logistic regression modeling? Why or why not?

(e) Finally, assess whether or not the proportional odds assumption holds for the model including both main effects of age and sex (but not their interaction). Based on the results of this analysis, what would be your recommendations for model choices if you wanted to include continuous age in the modeling? Would you recommend using ordinal or multinomial logistic regression? Is there anything else you might recommend?