BST 210 HOMEWORK #7

Due 11:59pm, Tuesday, November 19, 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(1a-1e, and 2a-2d), 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.

A study was performed to identify risk factors associated with giving birth to a low birth weight baby (< 2500 grams). More information is included in the "background" file. To control for the important factor of mother's age, one "case" (with low birth weight) was matched to three "controls" (without low birth weight), matching on mother's age. STRATUM is the matching variable, with LOW as the outcome variable. Our analysis will focus on looking at the (linear) effects of other factors, namely LWT, SMOKE, HT, UI, and PTD to predict low birth weight, appropriately adjusting for matching using conditional logistic regression. In fact, it turns out that if you were to use backward elimination with a 0.15 significance cutoff for conditional logistic model using all three controls matched with each case, you would find that the only significant predictor was PTD.

(a) Write a one or two sentence summary of the results for the conditional logistic model with PTD as the predictor, conditioning on the matched strata.

(b) One of your collaborators suggests that because mother's age is an important confounding variable, it should be included as a covariate. Run the model from part (a), but adding in AGE as a covariate. What do you find? Does that make sense? Briefly explain.

(c) Another investigator suggests that because integer mother's age was the only factor matched on originally, one could instead use AGE (rather than STRATUM) as the matching variable. Do you agree or not? Briefly explain. And, if you did that, how do your results change? Which do you prefer (and why)?

(d) Using a conditional logistic regression model with AGE as the matching variable, can we assess whether or not age is an effect modifier of the effect of PTD? If so, assess potential effect modification, or if not, briefly explain why not.

(e) Finally, compare the results of your conditional logistic regression to an (unconditional) logistic regression on the whole sample, adjusting for age as a covariate. Is this a good approach to use? Why or why not?

Problem 2.

A large study was performed looking at the dose-response effects of cigarette smoking on lung cancer incidence in British male physicians. The data to be analyzed were presented in Frome (*Biometrics*, 1983) and originally were collected by Doll and Hill. The data are given in the “fromelungcancer” file. We’ll be fitting a variety of models and making model comparisons.

(a) Our main interest is in the effects of CIGPDAY on lung cancer incidence. Do we have any evidence that (linear) SMOKEDUR is a confounder or an effect modifier of the effects of (linear) CIGPDAY on (log of) lung cancer incidence? (SMOKEDUR is a surrogate for both age and smoking duration, and is coded as age - 20, under the assumption that most smokers started smoking around age 20). Justify your responses, and summarize your overall findings briefly (e.g., in terms of incidence rate ratios).

Because of potential evidence of lack of fit and to consider possible nonlinear effects of CIGPDAY and SMOKEDUR on lung cancer incidence, a number of additional analyses should be performed.

(b) Consider a model including linear and quadratic effects of both CIGPDAY and SMOKEDUR. Does this model show improvements relative to the model including only linear covariates? Using this model, calculate a point estimate and 95% CI for the IRR for the effects of 20 vs. 0 cigarettes/day, and for the effects of 40 vs. 20 cigarettes/day, adjusting for linear and quadratic SMOKEDUR. Note that these point estimates and confidence intervals are not the same due to the quadratic effects of CIGPDAY included in this model.

(c) Because quadratic effects seem to be statistically significant, we might also want to run models that were even more complicated than quadratic. Given the small number of (effectively categorical) CIGPDAY and SMOKEDUR values, using generalized additive models or restricted cubic splines does not seem appealing. Those methods are more effective when you have a truly continuous covariate. Instead, fit a model with categorical CIGPDAY and SMOKEDUR, but no interaction. Using this model, calculate a point estimate and 95% CI for the IRR for the effects of 20.4 vs. 0 cigarettes/day, and for the effects of 40.8 vs. 20.4 cigarettes/day, adjusting for categorical SMOKEDUR. Note that these point estimates and confidence intervals are not the same due to the categorical (rather than linear) effects of CIGPDAY included in this model.

(d) Briefly, which model do you prefer and why?