Biostatistics 515/518 Winter 2020
HW 7 (9 questions – 8 required, Q6 is optional)

The class website contains a dataset about prisoner behavior in the California Department of Corrections (CDC). The CDC is interested whether its classification score for prisoners entering the system has predictive value for whether a prisoner will cause difficulty during his incarceration.

Note that the CDC makes assignment to maximum security partly based on its own classification score. The variable "TREAT" indicates whether a subject was assigned to a maximum security prison. The rules are much stricter in a maximum security prisons, and the prison has much greater control over the prisoners. Accordingly, prisoners in maximum security have fewer opportunities for bad behavior.

The CDC is interested in the inherent value of their classification score, as it would like to know whether the score is a good tool for identifying problem prisoners. In other words, the CDC is interested in the predictive value of their classification score aside from the effect of security level.

1. Draw a "causal diagram" to represent the relationships between the variables 'SCORE,' 'TREAT,' and 'Misbehavior.'

2. Perform a simple logistic regression directed towards the CDC's question of interest. Provide scientific interpretation of the slope. What is the estimated odds ratio and is there evidence that the classification score is associated with prisoner misconduct?

3. Further interpret your results: what does your model estimate to be the probability of misconduct for prisoners with scores 25, 50, and 75, respectively? You are not required to provide confidence intervals (but you are welcome to).

4. Based on the information provided, is it appropriate to adjust for 'TREAT'? Why or why not?

5. a. Regardless of your answer in (4), adjust the analysis you did in (2) for whether or not a prisoner was assigned to maximum security and summarize your results. Use language appropriate for a published paper.
b. According to the fitted logistic regression model with SCORE and TREAT as predictors, what is the probability of misconduct for prisoners with scores 25, 50, and 75? (You will need to answer this question for both levels of TREAT.)

6.(optional) Make an ROC curve for the predicted risks of misconduct for the model that only includes SCORE, and for the model that includes both TREAT and SCORE. Comment on each model's predictive capacity.

7. Assess the evidence that assignment to maximum security modifies the association between classification score and misconduct.

8. Regardless of whether you found statistically significant evidence in (7), interpret the logistic model that includes this interaction:
(a) Using the fitted model, give the estimated odds ratio (with CI) that characterizes the association between classification score and misconduct for prisoners **not** assigned to maximum security.
(b) Using the fitted model, give the estimated odds ratio (with CI) that characterizes the association between classification score and misconduct for prisoners assigned to maximum security.
(c) Speculate on the reasons for the different odds ratios you found in (a) and (b).

9. Return to the birth data used in discussion section. Create a variable using bwt that indicates a baby born less then 2500 grams. Consider the possible association between mother's age and having a low birthweight baby. Create a data frame containing data only from mothers who report no smoking or drinking alcohol during pregnancy and first birth (parity==0 & smoker=="N" & drinker=="N").
(a) For the question of interest, what role does mother's education play?
(b) For the question of interest, what role does the baby's sex play?
(c) Perform a logistic regression analysis with low-birthweight as the outcome, mother's age as the predictor of interest, and adjusting for race, mother's education, and the baby's sex. Interpret results using language suitable for a scientific publication.