NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp

Generalized Linear Models

Question #1: Birdkeeping & Lung Cancer

Load the `Sleuth2` library and extract the `case2002` dataset. This dataset reports results of a survey conducted from 1972 to 1981 in the Netherlands aiming to see if birdkeeping is a risk factor for lung cancer. Variables include whether or not an individual had lung cancer, whether or not they were birdkeeping, their gender, socioeconomic status, age, years of smoking, and average rate of smoking.

1. Perform some basic numerical and graphical EDA. In particular, comment on the scatterplots of the continuous variables colored by whether or not an individual had lung cancer. What might be good? What might be bad?
2. Fit a logistic regression predicting whether or not an individual has lung cancer that includes all variables in the model.
3. Briefly assess the appropriate residual plot and an influence plot for the model created in part 2.
4. Conduct and interpret an overall goodness of fit test for the model created in part 2.
5. Interpret the coefficient of gender on the log odds scale.
6. Interpret the coefficient of socioeconomic status on the odds scale.
7. Interpret the 95% confidence interval based on standard errors for the birdkeeping indicator on the log odds scale.
8. Interpret the 95% confidence interval based on standard errors for the years of smoking variable on the odds scale.
9. Fit a logistic regression predicting whether or not an individual has lung cancer that includes all variables in the model except the birdkeeping indicator.
10. Conduct and interpret an overall goodness of fit test for the model created in part 9.

11. Conduct and interpret a drop in deviance test comparing the two models you've created thus far. Which would you keep in favor of the other?
12. Fit a logistic regression predicting whether or not an individual has lung cancer based only on whether or not they have birds and the number of years they have been smoking.
13. Conduct and interpret a drop in deviance test comparing the model you created in part 12 to the model you created in part 2. Which would you keep in favor of the other?
14. Compare the models across:
    a. AIC
    b. BIC
    c. R^2_{dev}
    d. Give an argument for choosing the model created in part 12.
15. Using the model created in part 12, predict:
    a. The probability of having lung cancer for an individual with an average number of years smoking with and without birds within their household.
    b. The probability of having lung cancer for an individual with no years prior smoking with and without birds within their household.
16. Use the model created in part 12 to classify the observations in your dataset as having or not having lung cancer. Comment on how well the model performs as compared to the baseline.