Tam Nguyen

Hw4

25 October, 2018

"Substance Use Among College Athletes:

A Comparison Based on Sport/Team Affiliation"

1. Is use of the logistic regression model appropriate in this context?
    - Yes, since he is trying to determine which sports/teams are at the greatest risk for substance use. He has a binary response variable, so using logistic regression model is appropriate in this context.
2. What are your thoughts about whether the assumptions of the model are met?
    - Linearity: Checked by other researches. Also, his explanatory variables are binary and they are always linearity in this model.
    - Randomness: Yes, since the data was collected randomly by researches from CAS
    - Independence: Not really since people in the same team will affect the amount of drinking of other members.
3. Is the sampling method valid, and is the sample representative of the larger population about which he makes conclusions?
    - Sampling method: The dataset comes from Harvard School of Public Health College Alcohol Study (CAS). Researchers in CAS (1993) use a nested random sampling method to survey students at almost 200 US colleges and universities. All colleges and university in the dataset are generated from a list of 4-year

schools provided by the American Council on Education. Information of students involved in the researches is collected by their college administrators. The research uses the same sample of schools in 1997, 1999, and 2001 with follow-up surveys by year. I believe this is a good sample method to ensure the randomness of the dataset.

- The author uses data from the 1999 survey, which was the most current data available to the public. Also, this dataset contains information about more than 14000 students from 119 4-year schools across 39 states. With the variety in the colleges sample, the dataset from 1999 survey is a good representative of US colleges and universities.
- However, after cleaning the dataset by limiting the analysis to only respondents who participated in intercollegiate athletics, the sample size reduced to 2,316. This sample size is too small compared to the amount of US college students. Hence, the dataset that is used in the final model is not a good representative of US colleges and universities.

4. Are his conclusions valid for the tests he performs? Does he interpret things (like p-values, ORs, and CIs) correctly? Do you feel there are any limitations to this study that he didn't explicitly discuss?
    - He did a good job in interpreting ORs. However, he didn't analyze the CIs even when he included the CIs and ORs tables in the paper.
    - When interpreting the odds ratio, he used the term "more likely" a lot, which creates a sense of probability instead of the odds ratio.

- Independence problem: The independence of each observation in the dataset has some problems. People in the same team will affect the amount of drinking of other members, hence the independence is not guaranteed.
- Another question is: Does the model have enough data?
    - The final sample size is small. His model shows that female soccer athletes and male hockey athletes are two significant binge drink groups. However, when looking at the amount of observations in each group, soccer group has 336 observations, which is more than 4 times higher than the amount of observations for the hockey team.