Chapter 6 Part 3

05/10/2021

Case Study: Trying to Lose Weight

- The final case study uses individual-specific information so that our response, rather than the number of successes out of some number of trials, is simply a binary variable taking on values of 0 or 1 (for failure/success, no/yes, etc.).
- This type of problem—**binary logistic regression**—is exceedingly common in practice.

Case Study: Trying to Lose Weight

- Read the first few paragraphs HERE
- Stop before "Data organziation"

Data

- Sample of 500 teens from data collected in 2009 through the U.S. Youth Risk Behavior Surveillance System (YRBSS) [YRBS2009].
- The YRBSS is an annual national school-based survey conducted by the Centers for Disease Control and Prevention (CDC) and state, territorial, and local education and health agencies and tribal governments. More information on this survey can be found here.

Sample Questions

Here are the three questions from the YRBSS we use for our investigation:

- Q66. Which of the following are you trying to do about your weight?
- A. Lose weight
- B. Gain weight
- C. Stay the same weight
- D. I am not trying to do anything about my weight

Sample Questions

- Q81. On an average school day, how many hours do you watch TV?
- A. I do not watch TV on an average school day
- B. Less than 1 hour per day
- C. 1 hour per day
- D. 2 hours per day
- E. 3 hours per day
- F. 4 hours per day
- G. 5 or more hours per day

Sample Questions

- Q84. During the past 12 months, on how many sports teams did you play? (Include any teams run by your school or community groups.)
- A. 0 teams
- B. 1 team
- C. 2 teams
- D. 3 or more teams

Our Data

- Answers to Q66 are used to define our response variable: $Y = 1$ corresponds to "(A) trying to lose weight", while $Y = 0$ corresponds to the other non-missing values.
- Q84 provides information on students' sports participation and is treated as numerical, 0 through 3, with 3 representing 3 or more.
- As a proxy for media exposure, we use answers to Q81 as numerical values 0, 0.5, 1, 2, 3, 4, and 5, with 5 representing 5 or more.
- Media exposure and sports participation are also considered as categorical factors, that is, as variables with distinct levels which can be denoted by indicator variables as opposed to their numerical values.

Our Data

- BMI is included in this study as the percentile for a given BMI for members of the same sex.
- This facilitates comparisons when modeling with males and females.
- The terms *BMI* and *BMI percentile* will be used interchangeably with the understanding that we are always referring to the percentile.
- With our sample, we use only the cases that include all of the data for these four questions. This is referred to as a **complete case analysis**.
- That brings our sample of 500 to 445. There are limitations of complete case analyses that we address in the Discussion.

EDA Start

- Let us start our EDA in R.
- First we will go over the "DataPrepExampleWeightLoss" script, no need to copy, you will have the script. Also no need to run it yourself.
  - Afterword, read in data, called "risk2009.csv"
- Now the EDA analysis
  - Make a female column to be a 1 if female and 0 else
  - Look at lots of proportion tables
  - Stacked relative frequency plot split by weight loss.wt
  - Table broken down by sex, weight, and loss status

Summarizing some of the EDA

- Nearly half (44.7%) of our sample of 445 youths report that they are trying to lose weight, 48.1% of the sample are females, and 59.3% play on one or more sports teams.
- 8.8% report that they do not watch any TV on school days, whereas another 13.0% watched 5 or more hours each day.
- The median BMI percentile for our 445 youths is 68.
- The most dramatic difference in the proportions of those who are trying to lose weight is by sex; 58% of the females want to lose weight in contrast to only 32% of the males. This provides strong support for the inclusion of a sex term in every model considered.

Summary Table

- The summary table displays the mean BMI of those wanting and not wanting to lose weight for males and females.
  - The mean BMI is greater for those trying to lose weight compared to those not trying to lose weight, regardless of sex.
  - The size of the difference is remarkably similar for the two sexes.

Preparing for Logistic

- If we consider including a BMI term in our model(s), the logit should be linearly related to BMI.
- How do we check if this holds true?

Preparing for Logistic

- We can investigate this assumption by constructing an empirical logit plot.
- In order to calculate empirical logits, we first divide our data by sex.
- Within each sex, we generate 10 groups of equal sizes, the first holding the bottom 10% in BMI percentile for that sex, the second holding the next lowest 10%, etc.
- Within each group, we calculate the proportion, $\hat{p}$ that reported wanting to lose weight, and then the empirical log odds, $log(\frac{\hat{p}}{1-\hat{p}})$, that a young person in that group wants to lose weight.
- Let's create this in R

Reviewing the Plots

- WE can see the empirical logits for the BMI intervals by sex.
- Both males and females exhibit an increasing linear trend on the logit scale indicating that increasing BMI is associated with a greater desire to lose weight and that modeling log odds as a linear function of BMI is reasonable.
- The slope for the females appears to be similar to the slope for males, so we do not need to consider an interaction term between BMI and sex in the model. (We can still try it)

Stacked Bar Charts

- Next lets create two stacked bar charts, split by sex and split by loss.wt

Interpreting our Bar Chart

- Out of those who play sports, 44% want to lose weight, whereas 46% want to lose weight among those who do not play sports.
- We can use this plot to compare the proportion of respondents who want to lose weight by their sex and sport participation.
- The data suggest that sports participation is associated with the same or even a slightly lower desire to lose weight, contrary to what had originally been hypothesized.
- While the overall levels of those wanting to lose weight differ considerably between the sexes, the differences between those in and out of sports within sex appear to be very small.
- A term for sports participation or number of teams will be considered, but there is not compelling evidence that an interaction term will be needed.

Media Bar Chart and Another Empirical Logit

- Create relative frequency bar chart by sex and lose.wt of media
- Create an empirical logit of Media (it is continuous)

Media

- We saw that an increased exposure to media, here measured as hours of TV daily, is associated with increased desire to lose weight, particularly for females.
- Overall, the percentage who want to lose weight ranges from 38% of those watching 5 hours of TV per day to 55% among those watching 2 hours daily.
- There is minimal variation in the proportion wanting to lose weight with both sexes combined.
- We are more interested in differences between the sexes. We create empirical logits using the proportion of students trying to lose weight for each level of hours spent watching TV daily and look at the trends in the logits separately for males and females.
- There does not appear to be a linear relationship for males or females.

Initial Models

- Our strategy for modeling is to use our questions of interest and what we have learned in the exploratory data analysis.
- For each model we interpret the coefficient of interest, look at the corresponding Wald test and, as a final step, compare the deviances for the different models we considered.
- We first use a model where sex is our only predictor.
  - Lets fit this in R

Model 1

- Our estimated binomial regression model is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.75 + 1.09\text{female}$$

where $\hat{p}$ is the estimated proportion of youth wanting to lose weight.

- We can interpret the coefficient on `female` by exponentiating $e^{1.0919} = 2.98$ (95% CI $= (2.03, 4.41)$) indicating that the odds of a female trying to lose weight is nearly three times the odds of a male trying to lose weight ($Z = 5.520$, $p = 3.38e-08$).
- We retain sex in the model and consider adding the BMI percentile
- Fit a model now with female, bmipct

Model Discussion

- We see that there is statistically significant evidence ($Z = 8.997, p < .001$) that BMI is positively associated with the odds of trying to lose weight, after controlling for sex.
- This indicates that BMI percentile belongs in the model with sex.
- Our estimated binomial regression model is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -4.26 + 1.86\text{female} + 0.047\text{bmipct}$$

Model Discussion

- To interpret the coefficient on `bmipct`, we will consider a 10-unit increase in `bmipct`.
- Because $e^{10*0.047} = 1.602$, then there is an estimated 60.2% increase in the odds of wanting to lose weight for each additional 10 percentile points of BMI for members of the same sex.
- Just as we had done in other multiple regression models, we need to interpret our coefficient *given that the other variables remain constant*.
- An interaction term for BMI by sex was tested (not shown) and it was not significant ($Z = -0.70$, $p = 0.485$), so the effect of BMI does not differ by sex.

Model 3

- We next add `sport` to our model. Sports participation was considered for inclusion in the model in three ways: an indicator of sports participation (0 = no teams, 1 = one or more teams), treating the number of teams (0, 1, 2, or 3) as numeric, and treating the number of teams as a factor.
- The models we will fit will treat sports participation using an indicator variable, but all three models produced similar results. (not shown)
- Let us fit in R

Model Discussion

- Sports teams were not significant in any of these models, nor were interaction terms (sex by sports and bmipct by sports).
- As a result, sports participation was no longer considered for inclusion in the model.

Models Round 4

- We last look at adding `media` to our model.
- Let us do this in R.

Model Discussion

- Media is not a statistically significant term ($Z = -1.371$, $p = 0.170$).
- However, because our interest centers on how media may affect attempts to lose weight and how its effect might be different for females and males, we fit a model with a media term and a sex by media interaction term (not shown).
- Neither term was statistically significant, so we have no support in our data that media exposure as measured by hours spent watching TV is associated with the odds a teen is trying to lose weight after accounting for sex and BMI.

Drop-in-Deviance Tests

- Comparing models using differences in deviances requires that the models be **nested**, meaning each smaller model is a simplified version of the larger model.
- In our case, Models 1, 2, and 4 are nested, as are Models 1, 2, and 3, but Models 3 and 4 cannot be compared using a drop-in-deviance test.
- We also compare AIC

Test Discussion

- There is a large drop-in-deviance adding BMI to the model with sex (Model 1 to Model 2, 117.3), which is clearly statistically significant when compared to a χ^2 distribution with 1 df.
- The drop-in-deviance for adding an indicator variable for sports to the model with sex and BMI is only 462.99 - 462.59 $= 0.40$. There is a difference of a single parameter, so the drop-in-deviance would be compared to a χ^2 distribution with 1 df. The resulting *p*-value is very large (.53) suggesting that adding an indicator for sports is not helpful once we've already accounted for BMI and sex.

AIC Discussion

- For comparing Models 3 and 4, one approach is to look at the AIC. In this case, the AIC is (barely) smaller for the model with media, providing evidence that the latter model is slightly preferable.

Model Discussion and Summary

- We found that the odds of wanting to lose weight are considerably greater for females compared to males.
- In addition, respondents with greater BMI percentiles express a greater desire to lose weight for members of the same sex.
- Regardless of sex or BMI percentile, sports participation and TV watching are not associated with different odds for wanting to lose weight.

Limitations

- What do we think were limitations to this study?
- Sources of bias?

Model Discussion and Summary

- A limitation of this analysis is that we used complete cases in place of a method of *imputing* responses or modeling missingness.
- This reduced our sample from 500 to 445, and it may have introduced bias.
- For example, if respondents who watch a lot of TV were unwilling to reveal as much, and if they differed with respect to their desire to lose weight from those respondents who reported watching little TV, our inferences regarding the relationship between lots of TV and desire to lose weight may be biased.

Model Discussion and Summary

- Other limitations may result from definitions.
- Trying to lose weight is self-reported and may not correlate with any action undertaken to do so.
- The number of sports teams may not accurately reflect sports-related pressures to lose weight. For example, elite athletes may focus on a single sport and be subject to greater pressures, whereas athletes who casually participate in three sports may not feel any pressure to lose weight.
- Hours spent watching TV are not likely to encompass the totality of media exposure, particularly because exposure to celebrities occurs often online.

Model Discussion and Summary

- This analysis does not explore in any detail maladaptions—inappropriate motivations for wanting to lose weight.
- For example, we did not focus our study on subsets of respondents with low BMI who are attempting to lose weight.

Model Discussion and Summary

- It would be instructive to use data science methodologies to explore the entire data set of 16,000 instead of sampling 500.
- However, the types of exploration and models used here could translate to the larger sample size.
- Finally a limitation may be introduced as a result of the acknowledged variation in the administration of the YRBSS. States and local authorities are allowed to administer the survey as they see fit, which at times results in significant variation in sample selection and response.