GH 811 Problem Set 3

Background

You are a research fellow working at the Boston University School of Public Health. You have been asked to investigate the factors associated with cigarette initiation among never smoking youth in the United States. Fortunately, you have access to recent data from the Population Assessment of Tobacco and Health (PATH) Study which includes a wealth of information on tobacco use behaviors. To take advantage of the longitudinal design of the PATH Study, you decide to investigate baseline characteristics assessed at Wave 1 (2013-2014) and subsequent use of cigarettes after one year of follow-up (Wave 2, 2014-2015).

Source of Data

Population Assessment of Tobacco and Health (PATH) Study 2013-2015 (the condensed dataset and codebook are posted on the course website).

Additional information about the study including full data files and questionnaires can be found at https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/36231

Data Processing

Your analytic dataset should consist of youth who meet the following inclusion criteria:

- Reported that they had never smoked a cigarette at baseline

Output

1. How many respondents were eliminated as a result of applying the stated inclusion criteria and how many remain in the final analytic dataset?

2. Using a nested ifelse() statement, generate a new variable that combines the race category variable with the Hispanic ethnicity variable to produce the following categories: Non-Hispanic White, Non-Hispanic Black, Hispanic, Non-Hispanic Other.

3. What percentage of youth reported having ever tried a cigarette after one year of follow-up?

4. Generate descriptive statistics for your sample to go in Table 1. Please include the characteristics of the whole sample as well as characteristics of youth who try cigarettes at Wave 2 and youth who do not. Include all the following characteristics in your Table 1:

   a) Number and percent of youth in each age category
   b) Number and percent of youth who are male
   c) Number and percent of youth in each race/ethnicity category you created previously
   d) Number and percent of youth whose parent completed college or an advanced degree
   e) Number and percent of youth who live with a tobacco user
   f) Number and percent of youth who have tried alcohol
   g) Number and percent of youth who have ever used e-cigarettes

GH 811 Problem Set 3

5. Describe the pattern of characteristics for youth smokers versus non-smokers in Table 1.

6. Install ggplot2 and use this package to create some plots to visualize the prevalence of reporting having tried a cigarette at Wave 2:

    a. Create a bar chart showing the prevalence of ever cigarette use at Wave 2 by sex
    b. Create a bar chart showing the prevalence of ever cigarette use at Wave 2 by race/ethnicity
    c. Combine these plots into a single bar chart that shows prevalence of ever cigarette use by both sex and race/ethnicity

7. Investigate the association between living with a tobacco user and trying cigarettes at Wave 2 using a logistic regression. Report the odds ratio, 95% confidence intervals, and p-value. Interpret your results.

8. Use a multivariable logistic regression to identify predictors of cigarette initiation. Include all the variables from Table 1. Present results (odds ratios, 95% confidence intervals, p-values) in Table 2.

9. Describe the results in Table 2.

    a. Which baseline characteristics are predictors of cigarette initiation?
    b. Choose 3 significant predictors of cigarette initiation and interpret their odds ratios in sentence form

10. Congratulations on finishing your analysis! Now it is time to start thinking about limitations and next steps to discuss.

    a. Name 2-3 limitations of this analysis
    b. If you were to pursue this investigation further, what would be the next steps?

Submit answers to the above questions along with your tables, figures, and R code attached to the end of your word document by **2 pm April 10, 2018.**

BONUS: If you format your submission using R Markdown, we will give you 5 bonus points on your Problem Set!

GH 811 Problem Set 3

HINTS

Data processing- note that there are two variables for ever cigarette use. One variable if from Wave 1 (baseline) and the other variable is from Wave 2 (1 year of follow-up). Make sure you are picking the correct one!

1. Try performing an operation using the dim() function.

2. None

3. None

4. Your Table 1 should have 3 columns of data. Include a column with the characteristics of your entire analytic sample. To facilitate comparison between youth who try cigarettes and those who don’t, include a column with characteristics of youth who try cigarettes and a column for characteristics of youth who do not try cigarettes. Baseline characteristics are presented as Table 1 in most academic manuscripts, so you may want to try to find an example for formatting tips!

5. You do not need to describe every number presented in the data. Give us the most interesting highlights

6. Use of ggplot2 is required here. You should be producing 3 figures for this question. The as.factor() function may come in handy. Don’t forget to add labels, titles, and colors to your plots. You may also want to add bars representing 95% Confidence Intervals to your plots.

7. Again, the as.factor() function might be useful within your logistic regression. Note that the coefficients produced may express change in the log odds of the outcome for a shift to a different level of the predictor variable. To produce the odds ratios and their 95% CIs, you will need to complete additional steps of code. See R learning module on logistic regression for more detail.

8. See above comment

9. None

10. None

Bonus- R Markdown is a format for writing reproducible, dynamic reports with R. Use it to embed R code and results into slideshows, pdfs, html documents, Word files and more. Learn more at http://rmarkdown.rstudio.com/. This is what Professor Stokes uses to create his lecture slides!