Logistic Regression

Exploratory data analysis

Jenine Harris
Brown School

Exploratory data analysis for logistic regression

```
# import the libraries cleaned file
libraries <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/dat

# change data types
library(package = "tidyverse")
libraries.cleaned <- libraries %>%
  mutate(age = as.numeric(age))

# check the data
summary(object = libraries.cleaned)
```

```
##       age            sex               parent      disabled    uses.lib
##  Min.   :16.00   female:768   not parent:1205   no  :1340   no :809
##  1st Qu.:33.00   male  :833   parent    : 391   yes : 253   yes:792
##  Median :51.00                NA's      :   5   NA's:   8
##  Mean   :49.31
##  3rd Qu.:64.00
##  Max.   :95.00
##  NA's   :30
##      ses                      raceth                           educ
##  high  : 158   Hispanic          : 194   < HS                    :171
##  low   : 246   Non-Hispanic Black: 170   Four-year degree or more:658
##  medium:1197   Non-Hispanic White:1097   HS to 2-year degree     :772
##                NA's              : 140
```

Exploratory data analysis

- Before using `CreateTableOne()` to get descriptive statistics, check the distribution of any continuous variables to see if mean or median is more appropriate.

```
# examine the distribution of age
libraries.cleaned %>%
  ggplot(aes(x = age)) +
  geom_density(fill = "#7463AC", alpha = .6) +
  theme_minimal() +
  labs(y = "Probability density", x = "Age in years",
       title = "The distribution of age in the 2016 Pew Research\nCenter
```

Use CreateTableOne() to get descriptive stats

```r
# open tableone package
library(package = "tableone")

# get a table of descriptive statistics
table.desc <- CreateTableOne(data = libraries.cleaned)
print(table.desc, nonnormal = 'age', showAllLevels = TRUE)
```

```
##
##                       level                    Overall
##   n                                             1601
##   age (median [IQR])                           51.00 [33.00, 64.00]
##   sex (%)             female                     768 (48.0)
##                       male                       833 (52.0)
##   parent (%)          not parent                1205 (75.5)
##                       parent                     391 (24.5)
##   disabled (%)        no                        1340 (84.1)
##                       yes                        253 (15.9)
##   uses.lib (%)        no                         809 (50.5)
##                       yes                        792 (49.5)
##   ses (%)             high                       158 ( 9.9)
##                       low                        246 (15.4)
##                       medium                    1197 (74.8)
##   raceth (%)          Hispanic                   194 (13.3)
```

Using bivariate statistical tests prior to logistic

- One of the strategies used in some fields to develop a logistic regression model is to start with *bivariate* inferential tests for each of the potential predictors.

- Predictors that show a statistically significant relationship with the outcome are then entered into a larger model to see how they all work together to predict or explain the outcome of interest.

- In some cases, this could be considered a **questionable research practice** that could threaten research quality and reproducibility.

    - Questionable research practices (QRP) are strategies, like dropping (or adding) observations, that researchers use that introduce bias, typically in pursuit of statistical significance.

    - Using bivariate analyses is not always a QRP and is a good strategy for exploratory research.

    - However, since there was a lot of other research on library use already, this work isn't exploratory so use bivariate analyses as information but not for developing the statistical model.

Building a model of library use

- Based on prior research, age, sex, race-ethnicity, income, education, and rurality were important characteristics that relate to library use.

- Being a parent is a logical predictor of library use.

- Disabilities and library use might also be of interest.

- Rather than conducting separate bivariate statistical tests for each of these variables and library use, take advantage of the built-in statistical testing in the tableone package.

- `CreateTableOne()` can be used to create a table with descriptive statistics *and* bivariate statistical test results for any or all of the variables in a data frame.

  - The outcome of interest is library use, which is a categorical variable with two categories.

- Examining the relationship between this categorical variable and each of the other variables in the data set requires statistical tests to examine (1) the relationship between two categorical variables, and (2) the relationship between one binary categorical variable and a non-normally distributed continuous variable (age).

  - Chi-squared is useful for examining whether there was a statistically significant relationship between two categorical variables.

  - The Mann-Whitney U test works for examining the relationship between one categorical variable (with two categories) and one non-normal continuous one.

Creating the table

- `CreateTableOne()` automatically uses the appropriate test based on the data types.
- To make the table with columns representing the categories of a variable like `uses.lib`, the `strata =` argument can be used.
- When the `strata =` argument is used, the descriptive statistics in the table will be show for each variable for each category of the factor specified.
- In this case, using `strata = uses.lib` will result in descriptive statistics for the yes and no values of the `uses.lib` variable.
- In addition, when the `strata =` argument is used, the table shows the p-value association with a bivariate statistical test that is conducted as appropriate given the data types in the table.
- For variables in the table that are factor data types, this is chi-squared. For variables that are numeric data types, the test is one-way ANOVA, which is equivalent to an independent samples t-test when the means are compared across two groups.
- In the second function, `print()` there are a number of options for changing the table.
- One is to specify if any of the numeric variables do not meet the normality assumption for ANOVA; this is done with the `nonnormal =` option with the name of the variable that does not meet the normality assumption, like this `nonnormal = 'age'`.
- When non-normal is specified for a variable, the median and IQR are printed in the table and the Kruskal-Wallis test is used in lieu of ANOVA.

The table code

```
# get a table of descriptive statistics with bivariate tests
table.desc <- CreateTableOne(data = libraries.cleaned,
                             strata = 'uses.lib',
                             vars = c("age", "sex", "parent", "disabled"
                                      "ses", "raceth", "educ", "rurality

print(table.desc,
      nonnormal = 'age',
      showAllLevels = TRUE)
```

```
##                      Stratified by uses.lib
##                       level                    no
##   n                                              809
##   age (median [IQR])                           53.00 [35.00, 65.00]
##   sex (%)             female                     330 (40.8)
##                       male                       479 (59.2)
##   parent (%)          not parent                 639 (79.1)
##                       parent                     169 (20.9)
##   disabled (%)        no                         661 (82.0)
##                       yes                        145 (18.0)
##   ses (%)             high                        67 ( 8.3)
##                       low                        130 (16.1)
##                       medium                     612 (75.6)
##   raceth (%)          Hispanic                   111 (14.9)
##                       Non-Hispanic Black          79 (10.6)
##                       Non-Hispanic White         557 (74.6)
##   educ (%)            < HS                       102 (12.6)
##                       Four-year degree or more   276 (34.1)
```