Problem set 5: Diabetes, age, and weights (survey weights, not the other kind)

S470/670 Fall 2019

Do this problem set individually. Upload your submission to Canvas by 11:59 pm, Tuesday 29th October.

How does the proportion of U.S. residents with diabetes vary with age?

The data

The data is in the NHANES package. NHANES is a series of biennial surveys to study the health of the non-institutionalized civilian resident population of the United States. The survey does *not* take a simple random sample: it oversamples some groups to get more accurate estimates for some subpopulations. The NHANES package contains two data frames:

- NHANESraw contains data from the 2009–12 NHANES surveys.
- NHANES contains data resampled from the 2009–12 NHANES surveys, in order to resemble a simple random sample of U.S. residents.

In short, using the NHANESraw data frame requires taking the survey design into account—mostly straightforwardly by using survey weights—while using the NHANES data frame does not require weights.

The variables we require are:

- Diabetes: Has the participant been diagnosed with diabetes (Yes or No.) For the purpose of this analysis, ignore NA's.
- Age in years. Note that subjects older than 80 were recorded as 80.
- WTINT2YR (in NHANESraw only): survey weights. (There's another weighting variable, WTMEC2YR, that provides survey weights for only those participant on whom physical and health measurements were made, but we will not use it here.)

Questions

1. Omitting people whose diabetes status is NA,
   (a) Using the data frame NHANES, estimate the proportion of all U.S. residents with diabetes.
   (b) Using the data frame NHANESraw, find the proportion of people in the sample with diabetes (i.e. don't use weights.)

(c) Using the data frame `NHANESraw`, estimate the proportion of all U.S. residents with diabetes.

Compare how close these numbers are.

2. (a) Fit a logistic regression on the `NHANES` data to model the proportion of U.S. residents with diabetes by age, using age as a continuous predictor.

   (b) Fit a weighted logistic regression on the `NHANESraw` data to model the proportion of U.S. residents with diabetes by age, using age as a continuous predictor.

   Give an interpretation of the slope coefficients for these model for someone who doesn't know what logistic regression is. How close are the two models' coefficients to each other?

3. Calculate the weighted proportion of people who have diabetes for each age from 1 to 80 (again, ignoring NA's) in the `NHANESraw` data frame. Plot this proportion against age and add the curve from your weighted logistic regression to the plot. Describe ways in which your curve does and doesn't fit the data.

What to submit

Submit a PDF or HTML document with your write-up and a Markdown or .R file with your code.