Milestone 2

Martha Wetzel

3/26/2018

# Overview

For this assignment, I downloaded and cleaned five years of data from the Behavioral Risk Factor Surveillance System (BRFSS) survey. The data was downloaded from the following locations:

* [2015](https://www.cdc.gov/brfss/annual_data/2015/files/LLCP2015XPT.zip)
* [2014](https://www.cdc.gov/brfss/annual_data/2014/files/LLCP2014XPT.zip)
* [2013](https://www.cdc.gov/brfss/annual_data/2013/files/LLCP2013XPT.zip)
* [2012](https://www.cdc.gov/brfss/annual_data/2012/files/LLCP2012XPT.zip)
* [2011](https://www.cdc.gov/brfss/annual_data/2011/files/LLCP2011XPT.zip)

Code for this project can be found on [GitHub](https://github.com/mpwetzel1/Policies-and-Pain).

# Data Cleaning

The variables kept for the analysis are shown below in Table 1. Variables that had different names in different years were cleaned and standardized.

VarsByYear <- read\_csv("/Users/marthawetzel/Desktop/Emory/04\_Spring 2018/NRSG 741/Final Project/Variables By Year\_2010-2015.csv")

## Parsed with column specification:  
## cols(  
## `2011` = col\_character(),  
## `2012` = col\_character(),  
## `2013` = col\_character(),  
## `2014` = col\_character(),  
## `2015` = col\_character()  
## )

VarsByYear

## # A tibble: 20 x 5  
## `2011` `2012` `2013` `2014` `2015`   
## <chr> <chr> <chr> <chr> <chr>   
## 1 HIVRISK3 HIVRISK3 N/A N/A N/A   
## 2 LSATISFY LSATISFY LSATISFY LSATISFY LSATISFY  
## 3 QLACTLM2 QLACTLM2 QLACTLM2 QLACTLM2 QLACTLM2  
## 4 POORHLTH POORHLTH POORHLTH POORHLTH POORHLTH  
## 5 PHYSHLTH PHYSHLTH PHYSHLTH PHYSHLTH PHYSHLTH  
## 6 GENHLTH GENHLTH GENHLTH GENHLTH GENHLTH   
## 7 NUMADULT NUMADULT NUMADULT NUMADULT NUMADULT  
## 8 CHILDREN CHILDREN CHILDREN CHILDREN CHILDREN  
## 9 MEDCOST MEDCOST MEDCOST MEDCOST MEDCOST   
## 10 SMOKE100 SMOKE100 SMOKE100 SMOKE100 SMOKE100  
## 11 AGE AGE AGE80 AGE80 AGE80   
## 12 HISPANC2 HISPANC2 \_HISPANC \_HISPANC \_HISPANC  
## 13 \_PRACE \_PRACE \_PRACE1 \_PRACE1 \_PRACE1   
## 14 EDUCA EDUCA EDUCA EDUCA EDUCA   
## 15 EMPLOY EMPLOY EMPLOY1 EMPLOY1 EMPLOY1   
## 16 INCOME2 INCOME2 INCOME2 INCOME2 INCOME2   
## 17 CELLFON2 CELLFON2 CELLFON2 CELLFON2 CELLFON2  
## 18 <NA> N/A DECIDE DECIDE DECIDE   
## 19 IMONTH IMONTH IMONTH IMONTH IMONTH   
## 20 IYEAR IYEAR IYEAR IYEAR IYEAR

After careful review of the data dictionaries, it became clear that several of my secondary analyses would not be possible. In particular, it will not be possible to examine HIV risk behavior or sleeplessness. In the case of HIV risk behavior, the question was not asked consistently and the wording was changed substantially between years, making the responses incomparable. For sleeplessness, the question was not consistently.

## Workflow

The following broad steps were performed in the data import/cleaning process:

1. Survey data acquisition
   * Download data dictionary
   * Download and unzip data files
   * Import the SAS Transport formatted files
   * Clear the raw files to reduce harddrive space (raw files took up a gigabyte each)
   * Verify that file included the correct year(s)
2. Law data set
   * I had previously compiled a data set showing the implementation dates of PDMP and marijuana laws at the state level

* Imported the SAS file containing the laws
* Corrected the dates to have the R origin point instead of the SAS origin point

1. Combine data

* Stack years of BRFSS data
* Verify that the numbers of observations per year were similar
* Inner joing the BRFSS and law data on state and year

1. Prepare the focal predictors for analysis

* Create an approximate interview data based on month and year that interview was conducted
* Create flags indicating if the individual lived in a state that had an active policy at the time of the interview

1. Prepare the covariates and dependent variables for analysis

* Standardize variables that had different names or slightly different response options in different survey years
* Create binary variables to use in models.

### Assignment Directions

A working prototype of your project is due on Wednesday 15 March by 5:00 pm EDT. You will submit a PDF file. This document will contain two links; 1. a link to the Rmd file that created it in your GitHub repo, and 2. a link to the source of your original data. Please put your name on the first page of the document.

For this milestone we expect that you will have acquired, cleaned, and explored your dataset. You will document these activities in your files. You will also explain in detail the components of your final analysis. If you have deviated from your original plans, please describe what is different and the reasons that have led to this. You will explain your workflow, that is, how you went from acquiring your data to getting (close to) an answer to your question. Are there ancillary questions that have arisen as you have gone through this process? What are they?