Assignment 1

Descriptive analysis of a multilevel data set

Due February 20 on bCourses site

Introduction

In the Methods in Social Epidemiology course, you will be conducting analysis exercises using subset of variables from the New York Social Environment Study (NYSES) data set. The NYSES is a cross-sectional multilevel study designed to examine neighborhood level exposures, including economic, social and structural characteristics, and substance use in New York City (NYC). The NYSES was conducted between June and December of 2005. We used random-digit-dial methods to contact and interview 4000 NYC residents. One adult 18 years or older was interviewed by telephone in each household. The neighborhood units in the NYSES data set are the 59 community districts in NYC.

For the analysis exercises in the class, we will focus on neighborhood level socioeconomic status exposures as captured by median income and the percent living below poverty. We will focus on the outcome of binge drinking. Data on a variety of other demographic and socioeconomic characteristics (age, race, education, income) are provided as confounders.

Your first assignment is to familiarize yourself with this data set, including coding and formatting variables as you see fit, and conducting descriptive analyses. The idea is to imagine you have received these data (perhaps as part of a new GSR position, or perhaps in starting a thesis or dissertation analysis) and are interested ultimately in testing a hypothesis about the relation between neighborhood socioeconomic status and binge drinking. You would first want to familiarize yourself with the data set, make it user friendly, and get a sense of the relations between the variables of interest. Below are some examples of R code that you can adapt to describe this data set – the examples do not use the variables in this data set but should be easily modified. You can certainly go beyond the code below – the example code is provided as a starting point. You can use different statistical software (e.g., Stata or SAS) if you prefer, but all of the sample code for the assignments will be provided in R. Feel free to work in groups to do this assignment, however each person must write up their own version of the assignment and their interpretation of the results. The data documentation file describes the variables in the data set, and the .csv file contains the data.

Sample R Code

# load packages used in this assignment

# only need to install packages once

install.packages("dplyr")

install.packages("ggplot2")

install.packages("nnet")

# need to call the libraries every time you begin a new R session

library(dplyr)

library(ggplot2)

library(nnet)

# read in data – suppose the file dataset.csv contains continuous variables var1 and var2, and a binary variable outcome

df <- read.csv("c:/dataset.csv")

# create a categorical variable from a continuous variable

df$catvar1 <- df$var1

df$catvar1 <- ifelse(df$catvar1<=500,0,ifelse(df$catvar1>500,1,NA))

# describe variables

summary(df$var1)

table(df$var1)

hist(df$var1)

with(df, table(var1,var2))

with(df, table(var1,var2, exclude=NULL))

df %>% group\_by(catvar1) %>% summarise(mean\_outcome = mean(outcome))

# bivariable relations

# to do a Pearson's chi-squared test

x2 <- chisq.test(df$var1, df$var2)

# to see the results of the test

x2

# to see the table of observed

x2$observed

# to see the percents by row

prop.table(x2$observed, 1)

# to see the percents by column

prop.table(x2$observed, 2)

# calculate correlation between variables

cor.test(df$var1, df$var2)

# summarize relationship between variables in a plot with a lowess line

ggplot(df) + geom\_point(aes(x=var1, y=outcome)) +

geom\_line(aes(x=var1, y=predict(loess(outcome~var1)))) + theme\_bw()