Poster Project Report

George Rhodes

12/8/2017

Project Topic and Relevance

The Substance Abuse and Mental Health Services Administration (SAMHSA) claims that 21.6 million Americans age 12 and over struggled with a substance use disorder (SUD) in 2013. The global focus on health equity and the social injustice of health disparities highlight the need to explore social determinants of health. Since socioeconomic status (SES) influences the availability, utilization, and impact of health services, attempts to treat the addicted population must take SES into account (Turner, 2013). SES is negatively associated with morbidity and detrimental health behaviors such as smoking and physical inactivity (Pampel et al., 2010; Phalen et al, 2010). Substance use specifically has a more complex dynamic, as people of low SES report both higher rates of binge drinking and abstinence, while those of higher SES report higher frequency of light drinking (Cerdá et al., 2011). But what role does SES play on the ability to recover from addiction? Does SES impact the treatment outcomes of respondents to the 2015 National Survey on Drug Use and health (NSDUH) who are 18 years old and older?

Fundamental cause theory, developed by Link and Phalen, attributes the lasting relationship between morbidity and SES to "an array of resources, such as money, knowledge, prestige, power, and beneficial social connections that protect health" (Phalen et al., 2010). Through this lens, we can begin to understand the social dynamics of health disparities. Extending this theory to addiction, those of higher SES would be expected to incur fewer negative consequences and more easily recover from addiction.

Finding Data

Knowing my topic of interest with the goal of using this project to familiarize myself with the dataset I would like to use for my thesis project, I searched haphazardly for addiction statistics and agencies for a couple days before stumbling onto the NSDUH. The NSDUH is an annual survey of a nationally representative sample conducted by the SAMHSA within the U.S. Department of Health and Human Services. The exhaustive list of questions on substance use and mental health drew me to this dataset. I wasn't sure I'd find any quantitative data on both SES measures and treatment outcomes, but many operationalization options exist within this dataset, as well as general demographic data to control for as many variables as possible.

Originally I just wanted to find SES measures and ways to subdivide the sample by those who have experienced successful treatment outcomes and those that had not. Data of income, family income, and level of education worked for SES measures. And the combination of the following two questions allowed me to divide the sample by treatment outcomes: 1) Have you ever received alcohol or drug treatment? 2) During the past 12 months, was there a month or more when you spent a lot of your time getting or drinking alcohol?

2,956 respondents have never been to drug or alcohol treatment, and 11,935 respondents said they had not used alcohol in the past 12 months or used less than 6 days. The 11,935 was distinguished from the 15,536 who chose "never used alcohol," both of which remained consistent throughout the entire line of questions asking about alcohol use over the past 12 months. So the first question establishes who has been to treatment and the second establishes who has been abstinent fro the past 12 months (or drank less than 6 days). Determining that these questions could act as a proxy for distinguishing between those who have been to treatment and have changed their behavior and those that have attended treatment and not changed their behavior, I committing to cleaning and analyzing this dataset.

Data Structure & Data Munging

The original structure of the 2015 public use NDSUH was large and cumbersome for an inexperienced researcher such as myself. The 57,146 observations of 2,666 variables intimidated me. However, I later came to appreciate the polished nature and detailed codebook of the dataset. Now, after working with this data for 2 months, I feel very comfortable manipulating for analysis and creating visualizations with it.

First, I created a subset of the data with only the variables of interest using the pipe operator (%>%) and the select() function. Next, I had to rename them using the colnames() function to make coding easier. I chose variables (or columns) representing the respondent's identity; have they been to treatment; have they drank alcohol in the last 12 months; their age, sex, race, marital status, income, family income, education level, relationship to the poverty level, whether they reside in urban or rural tract, how many days they drank in the past 30, and how many days have they binge drank (4 for female, 5 for male) in the past 30. Initially, this was difficult and I needed to work out a few bugs, such as making all the original column names uniformly lowercase using the tolower() function as case sensitivity gave me issues.

Once I had all my data named appropriately, I needed to clean the data through a long trial and error process that continues whenever I try to explore a new relationship. This means taking the two questions about attending treatment and remaining abstinent for the past 12 months and creating a new variable "treated_sober" using the mutate() function and assigned the new dataframe to a new object. I imbedded an ifelse loop within the mutate to create three catagories: 1) Treated_sober = have attended treatment and not drank in last 12 months 2) Treated_drinking = have attended treatment and drank in last 12 months 3) No_treatment = have not attended treatment.

Originally, I named the no_treatment group "untreated," but this caused confusion when people attempted to interpret my visualizations and thought that the untreated group represented those who have untreated addictions. For the most part, the no_treatment group represents society as a whole, as most people have not been to drug and alcohol treatment. This could be because they are not problem drinkers, or maybe they are problem drinkers who have not received treatment.

Further cleaning required filtering to respondents 18 and older using the filter() function. This made the education level more meaningful, as it didn't lump everybody 12-18 in the "less than highschool" category. I also had to go in and rename catagories so they would show up in my visualizations. While I knew what the graphs meant, my audience could not interpret the values of 1-5 to signify level of education. I used the mapvalues() function to rename income, family income, and level of education according to the codebook.

The most difficult part was getting my visualizations to show both what I wanted, and something interesting. I imagined creating profiles of each SES measure across the three treated_sober subgroups. This required multiple layers of geom_bars, each one with a different subset of data for each category of treated_sober so that the proportions would be within the subgroup, not across the total population. Without this step, the proportions would be across the whole sample and the untreated group would be overrepresented and dominate the entire bar graph. I changed the color of each geom_bar, added in labels, facetted the geom_bars, used the economist theme, and then changed the angle of the x-axis labels so they were both legible and accurate.

Creating the mosaic plots and running chi-squared tests were much easier, as at that point, the data was clean and I had gained confidence in my coding. I even started running the visualizations and models on more and more variables. I created an multivariate of most of the variables, all accurately recoded, to see how they impacted number of days binge drinking in the month among those who have been to treatment. While no statistically significant associations came from this, every variable seemed to explain more of the variation as the adjusted R-squared increased. I chose only to display the mosaic plots of the variables education and treated_sober and the variables family income and treated_sober for the sake of significance and parsimony.

Reproducible Research

This process has taught me the value of principles of reproducibility. It is very clear to me how easy it is to get lost in the weeds, make code convoluted and difficult to read and understand, and to have all required pieces built into the environment. It took me a couple frustrated days before I realizedtime I open an Rmarkdown file I need to run everything in to have all objects and values saved in the environment. This was particularly frustrating when working on separate pieces of the data cleaning process as well as creating visualizations and then trying to piece them together.

By using packrat, loading all necessary libraries within the RMD files, pcushing everything to a public access Github repository, and providing supplemental documentation to explain decisions made along the way, I have attempted to make this research reproducible. Add onto that relatively clean code, useful variable names, only a few too many pipe operators (I struggled with larger imbedded chains), and lots of comments for each line, and I think that a researcher familiar with R could follow the logic of the process to reproduce it as well as critique it. If this research is not fully reproducible, at least I have learned enough to be able to walk somebody through the process step by step, explain it, and repeat it if necessary.