Predictors of Overall Health in the National Survey of Drug Use and Health

Jason E. Piccone, Ph.D.

Abstract

The following is a brief examination of predictors for overall health from the National Survey of Drug Use and Health (NSDUH). The NSDUH is a state-stratified sample of residents of the United States, age 12 and over, and was conducted by the United States Department of Health Services. The current analysis consists of 125 predictor variables and 10,000 adult respondents. SAS software is used to produce all analyses, including a least absolute shrinkage and selection operator (Lasso) regression for variable selection and prediction. Fifty-four variables are selected producing an initial benchmark model. In addition, the ten strongest predictor variables are explored to better understand their relationship with respondents' overall health.

Predictors of Overall Health in the National Survey of Drug Use and Health

Understanding the complex relationship between drug use and demographic variables with health is exceedingly important. The goal of this report is to explore these relationships and to demonstrate advanced statistical/machine learning techniques using SAS software. This analysis is based on the National Survey on Drug Use and Health (NSDUH), 2012 (ICPSR 34933).

Data Acquisition and Preparation

The NSDUH (2012) data can be downloaded from: http://www.icpsr.umich.edu/icpsrweb/content/samhda/dataportal.html. The survey was conducted through a stratified sample such that samples were taken in proportion to the population of each state. Responses were gathered in field interviews and respondents included individuals aged 12 and over across the United States. The raw data consisted of 55,268 respondents and included 3,119 predictive variables. This analysis, however, consists of a pre-cleaned paired down version of this dataset. The resultant data set consists of 125 predictor variables and the outcome variable of overall health. Continuous variables were standardized and categorical variables were dummy code when needed. Since the analyses to be conducted were performed on the University Edition of SAS, the dataset had to be reduced to 10,000 adult (18 and over) respondents.

Data Visualization

It is always important to visually inspect the data to ensure that the recoding process was successful and to better understand the nature of the data. Many of the predictor variables in this dataset are imbalanced. For example, very few people tried some of the many drugs included in the survey (such as crack cocaine). This imbalance should be considered, especially when the imbalance is great and in particular, when it involves the outcome variables. In this case, the outcome variable of interest is overall health. As Figure 1 below indicates, it is positively skewed. While not reported here, I performed the primary analysis after normalizing this distribution through transformations, but they did not improve the predictive efficacy of the model. We can also see, after viewing the basic demographics (Figure 2, a-d) that the sample is largely Caucasian, female, unmarried, and young.

Figure 1

Distribution of Overall Health

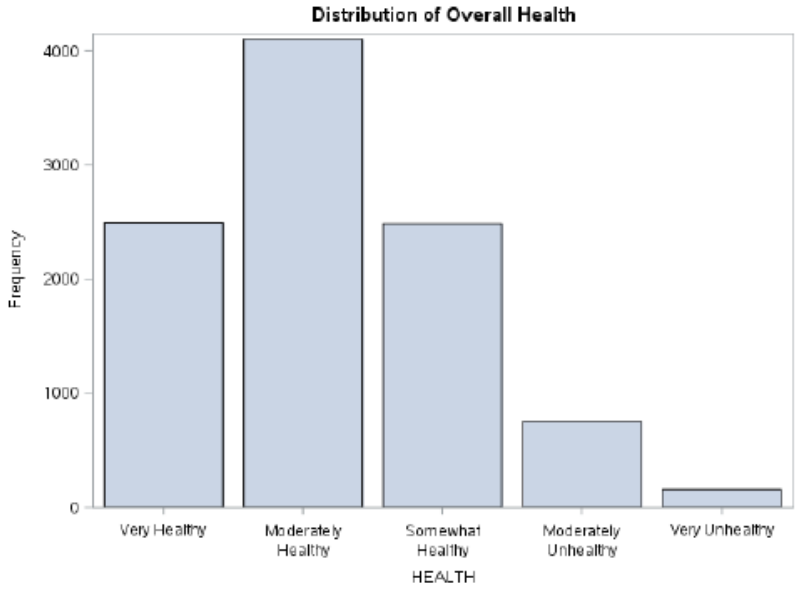
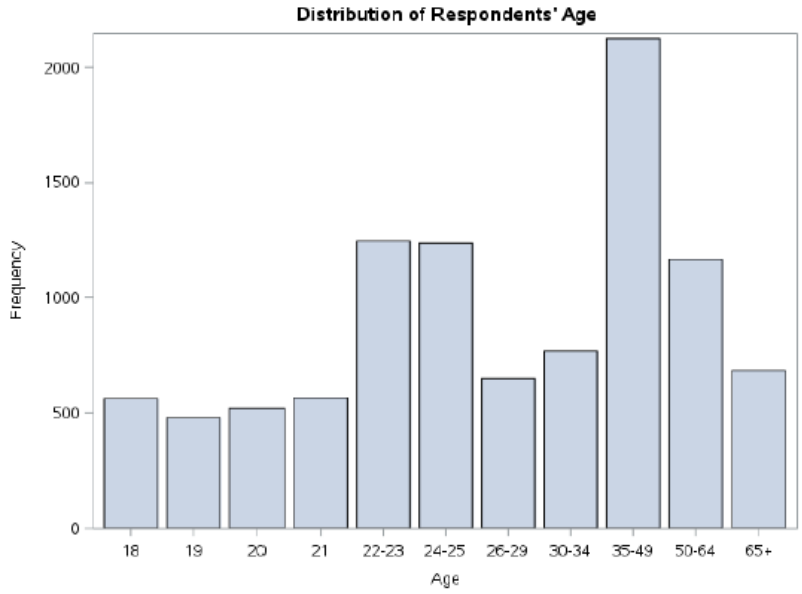
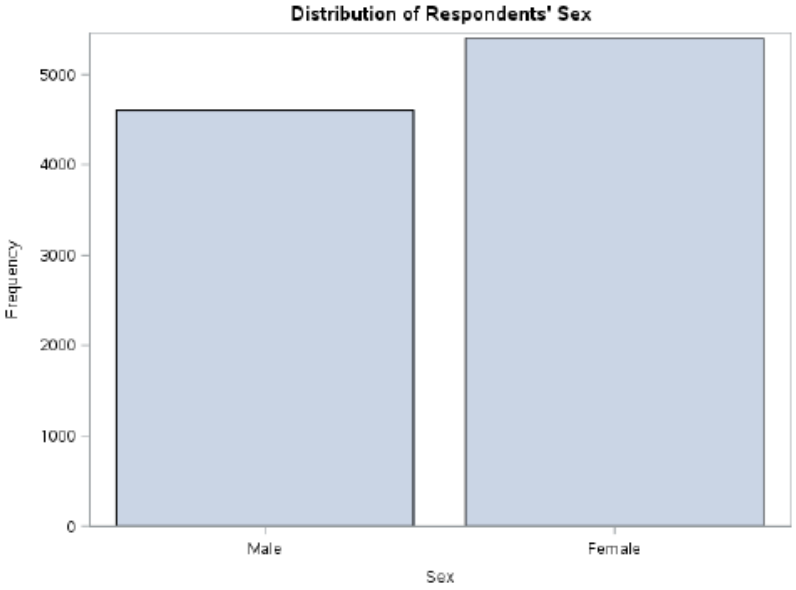
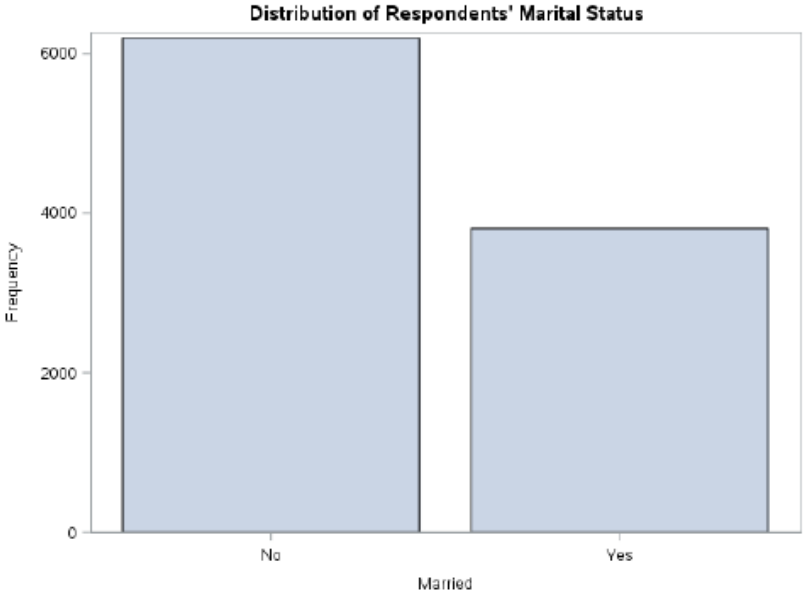


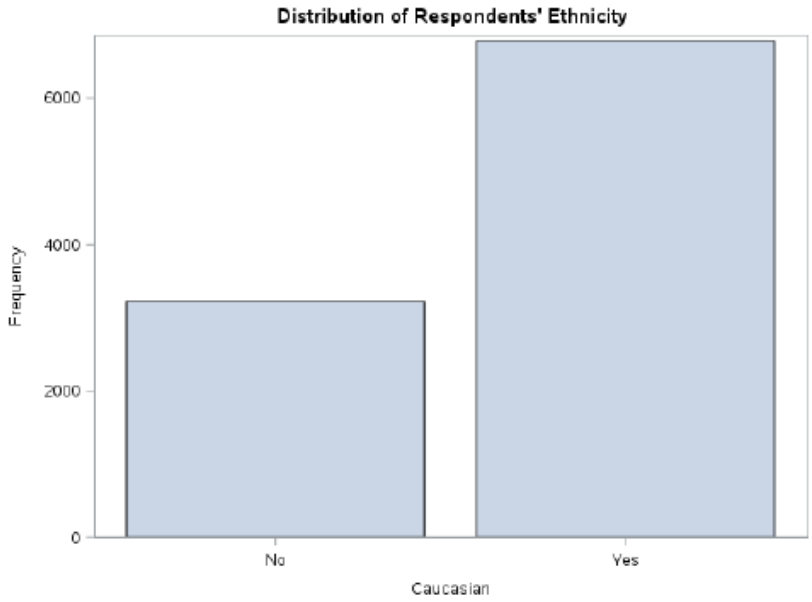
Figure 2 (a-d)

Distribution of Respondents' (a) Age, (b) Sex, (c) Marital Status, and (d) Ethnicity




Distribution of Respondents' Sex
5000
4000
3000
2000
1000
0
Frequency
Male
Female
Sex


Distribution of Respondents' Marital Status
6000
4000
2000
0
Frequency
No
Yes
Married



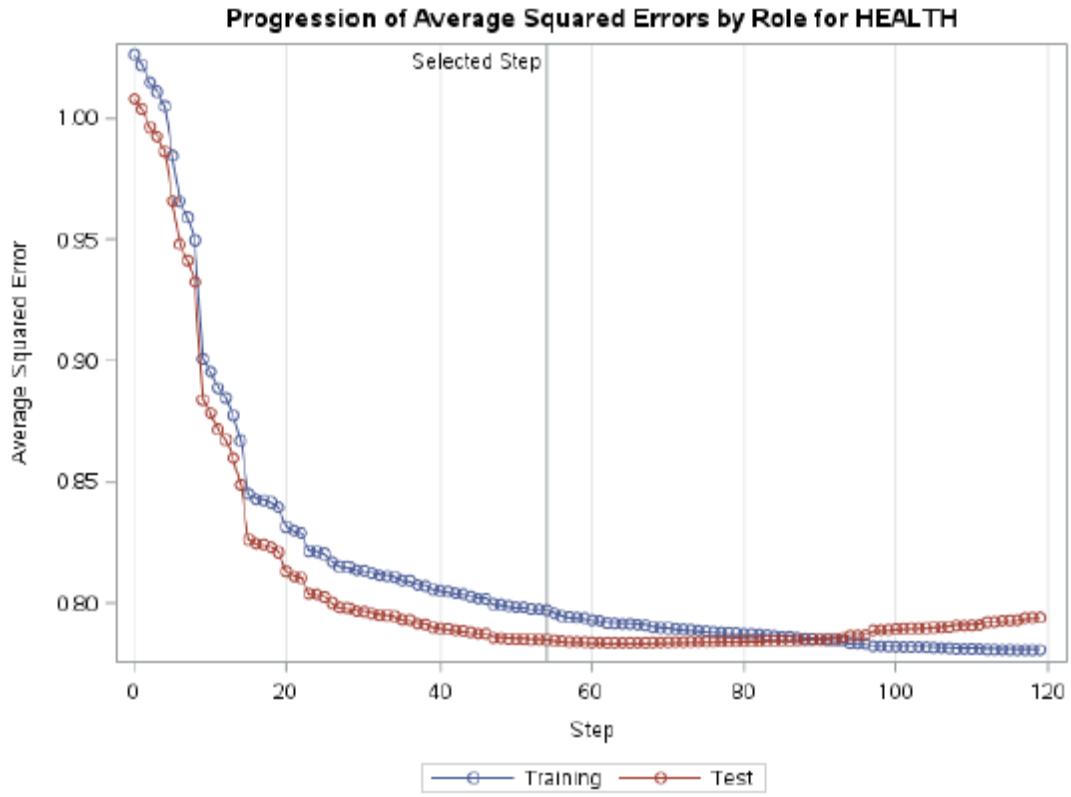
Predicting Overall Health

With the data cleaned, standardized, and visualized we can turn to the primary analysis of predicting overall health. Many powerful algorithms could handle such an analysis, but the least absolute shrinkage and selection operator (Lasso) regression is a strong choice. Lasso is like stepwise regression in that it selects a subset of variables to include in the final analysis. It is superior to stepwise regression because it also shrinks regression coefficients, which reduces overfitting. To additionally improve the model, the sample was split into 70/30 train and test samples and 10-fold cross validation was used. Traditional significance tests and estimates of effect size are not appropriate for such a model and interpretation should be conservative. It is appropriate, however, to report the mean squared error on the hold-out test set, which is 0.785. This is a solid benchmark, for which we can compare more refined models against later.

Figure 3 is useful to examine model efficacy as predictors are added to the train and test set. This example is highly unusual in that the test set outperforms the train set through most iterations. It is typical, however, in that the train set model continues to improve and the test set eventually gets worse as more variables are entered. The model chosen is the one that best balances the risk of over and underfitting, which in this case, is when 54 predictors are entered.

Figure 3

Average Squared Errors across Model Iteration for Train and Test Sets



Exploring the Top Predictors

Figure 4 is a truncated (top-10) list of the most statistically meaningful predictor variables. These variables, in order include (a) education level, (b) how often you feel sad or depressed, (c) how often you feel that everything is an effort, (d) how often you feel hopeless, (f) are you covered by Medicare, (g) are you covered by private insurance, (h) age, (i) how often you feel worthless, (j) number of times treated in the emergency room in the past 12 months, and (k) took prescription medication for a mental health condition in the past 12 months. ASE is the average square error, which is presented for the train and test sets, where the lower score is desirable. The CV press is the sum of the residual sum of scores for the test data set. If this figure was not truncated, you would see that after the 54th variable, the CV press begins to increase, indicating less optimal models.

Figure 4

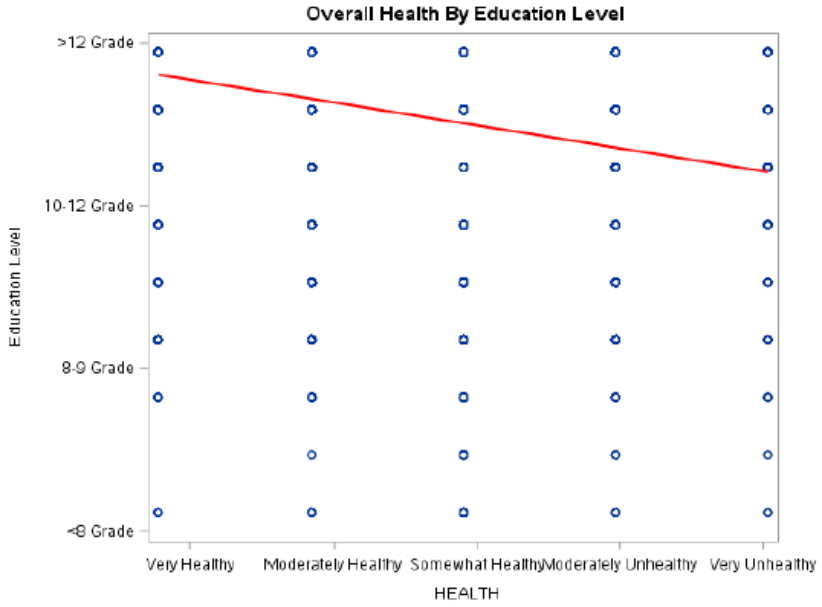
Truncated (Top-10) List of Predictor Variables

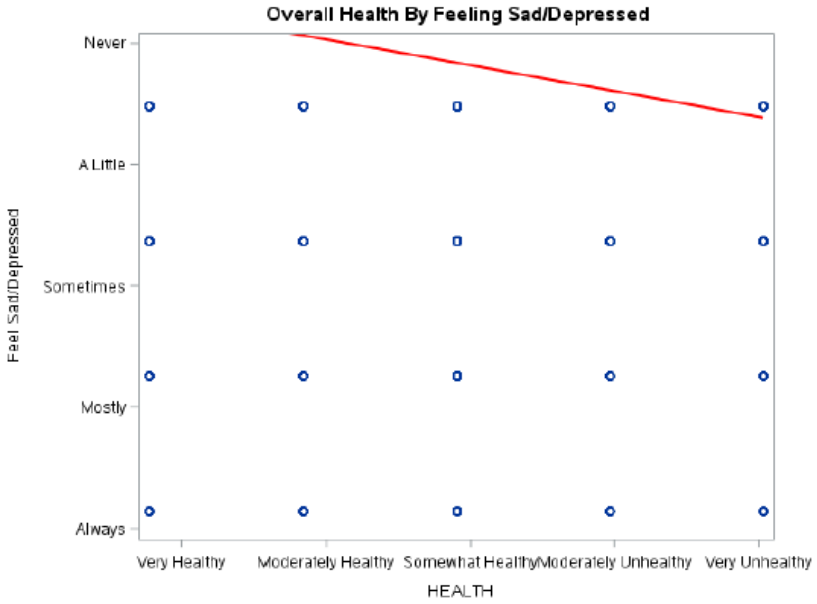
LAR Selection Summary					
Step	Effect Entered	Number Effects In	ASE	Test ASE	CV PRESS
0	Intercept	1	1.0262	1.0081	7187.6559
1	IREDUC2	2	1.0219	1.0036	6812.6145
2	DSTCHR30	3	1.0147	0.9963	6553.3489
3	DSTEFF30	4	1.0109	0.9925	6475.8032
4	DSTHOP30	5	1.0048	0.9864	6449.2853
5	MEDICARE	6	0.9845	0.9660	6203.9486
6	PRVHLTIN	7	0.9658	0.9479	6159.0960
7	AGE2	8	0.9590	0.9414	5939.6194
8	DSTNGD30	9	0.9496	0.9323	5932.0810
9	NMERTMT2	10	0.9006	0.8838	5865.4546
10	AURXYR	11	0.8954	0.8785	5842.8561

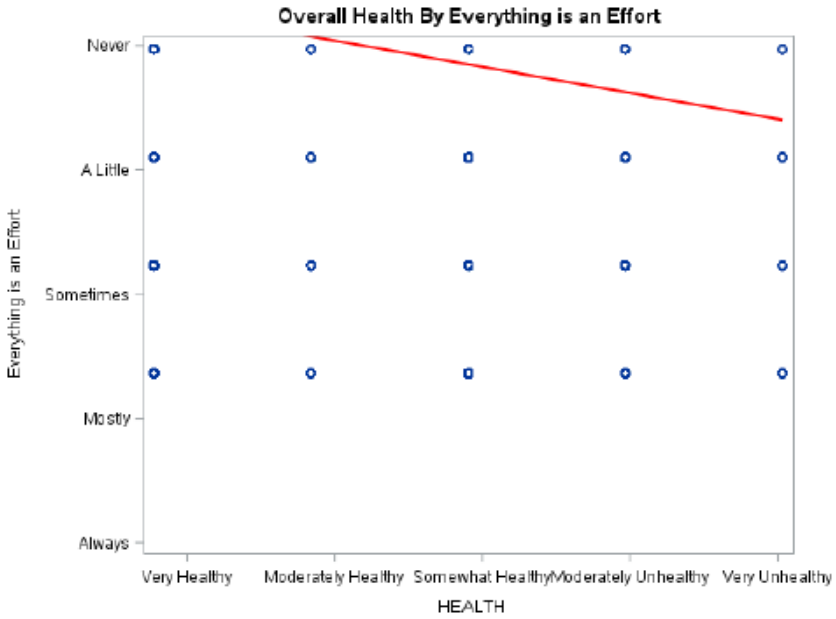
It is important to examine the variables included in the final model to determine that they make logical sense. For example, a predictor variable such as “I am sick most of the time” would be too similar to the predicted variable. Such a variable would account for much of the variability in the outcome, masking the importance of other, more relevant predictors. I created figures to better demonstrate the relationship between each predictor and overall health. For the sake of space, only the top four variables are displayed below (see Figure 5 (a-d)). You can see that better overall health is associated with more education, and less feelings of sadness, that everything is an effort, and hopelessness.

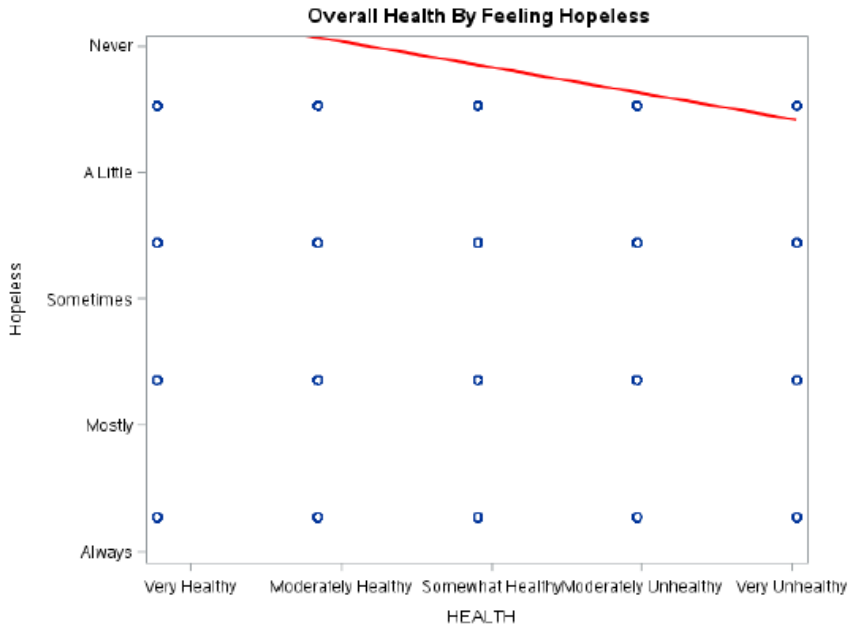
Figure 5

Relationship between Overall Health and (a) Education, (b) Depressed, (c) Effort, and (d) Hopeless









Conclusions

The preceding analysis presents a strong benchmark model predicting overall health using a Lasso regression. While the current model presents a fairly simple Lasso regression, it could be improved through the inclusion of interactions and polynomials. In addition, other successful algorithms could be explored, including random forests and gradient boosting. Nevertheless, this sets a nice benchmark model and highlights important predictive variables.

It is interesting that many of the top-10 predictors are variables that are situational and possibly controllable (such as education level, depression, and feelings of hopelessness). While obviously the nature of this research design does not allow for causal conclusions, work to improve the education

level and mental health of Americans could be associated with increased health. Further research, and in particular, randomized studies could better ascertain the causal connection between these variables, leading to a clearer course for intervention strategies.