Predictors of Stroke Prevalence in Milwaukee County

View R code at https://github.com/MichelleAndersonMSDS/Predictors-of-Stroke

Michelle Anderson

4/23/2023

Stroke is a cardiovascular disease that affects blood flow to the brain. A stroke occurs when a blood clot or blood vessel rupture obstructs the flow of blood to the brain. When the brain cannot get enough blood (and therefore oxygen) parts of the brain become damaged or die. Stroke is the 5th leading cause of death and a major cause of disability in the United States.

Improving stroke outcomes begins with pre-hospital care. In the field, emergency medical services (EMS) professionals, such as paramedics and emergency technicians, collect valuable clinical information about stroke symptoms and when they started prior to emergency department arrival, ensure patients are transported to a stroke-receiving hospital or certified stroke center as quickly as possible, and pre-alert hospital-based stroke teams of an arriving patient needing stroke evaluation and care.

In addition to acute care, EMS agencies are working to mitigate non-life threatening issues before they become an emergency though community mobile integrated health programs. Understanding where the greatest disease burden exists within a community and which community-level risk factors are present can help direct limited-resources to where they are needed most.

This study aims to explore the degree to which neighborhood-level social determinants of health (SDOH) can help predict crude prevalence of stroke by census tract. Examples of SDOH examined in this analysis include: employment, household income, educational attainment, language barriers, proximity to healthcare, access to health insurance, housing affordability, and transportation. Additional covariates include prevalence of related disease processes (e.g., hypertension, diabetes, etc.) and health risk behavior indicators (e.g., smoking, exercise, etc.).

A census tract-level health data set was created by pooling data from the 500 Cities Project, the PLACES: Local Data for Public Health project, and the Social Determinants of Health Database. The final data set includes 5 years of data for 212 census tracts in Wisconsin[1]. The median crude prevalence of ever being diagnosed with stroke by a physician among census tract residents who are 18 years or older was 3.5% (range: 0.8%-9.2%).

All variables in the final data set are numeric with no more than 4.9% of the data is missing on any given variable. Median imputation by census tract was applied to address missing values. The Box-Cox method was used to recommend reasonable transformations for heavily skewed variables (see Figure 1, Table 1). A small number of variables were combined to create new, more meaningful variables (e.g., variables for the percent of households with an income of \$0-\$10,000, \$10,001-\$14,999, and \$15,000-\$24,999 were collapsed to create one variable for the percent of households with income below \$25,000).

[1]*Note: The original 500 Cities Project included data for the following Wisconsin cities: Milwaukee, Madison, Green Bay, Kenosha, Racine, Appleton, and Waukesha. The PLACES project (which replaced the 500 Cities Project in 2020) includes data for all census tracts in Wisconsin. For consistency, this analysis was restricted to the census tracts used in the 500 Cities Project.

Figure 1. Square root transformation of the response variable (STROKE_CrudePrev)

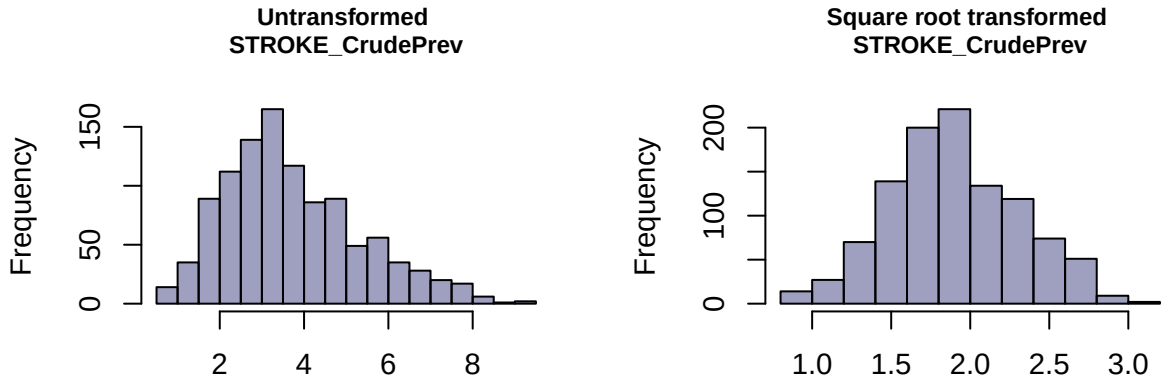
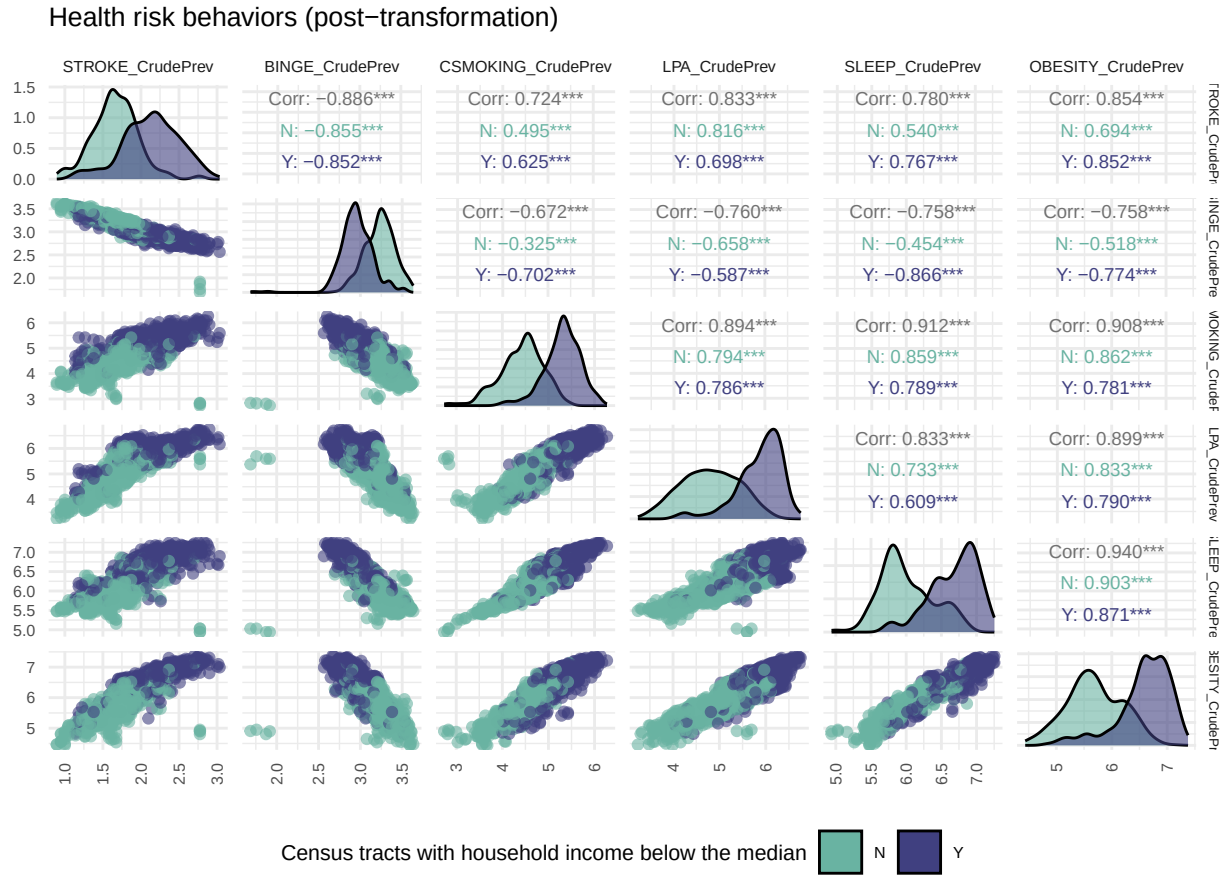


Table 1. Boxcox recommended transformations (sample)

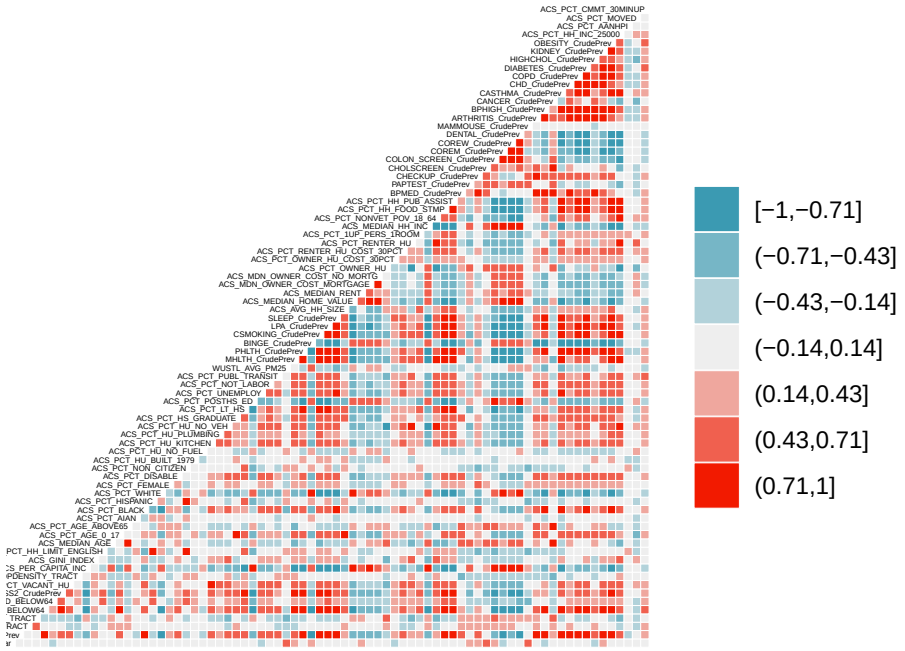
	Variable	Lambda	Recommended transformation
12	ACS_PCT_AGE_0_17	0.0	log(x)
13	ACS_PCT_AGE_ABOVE65	0.6	sqrt(x)
14	ACS_PCT_AIAN	0.7	sqrt(x)
15	ACS_PCT_BLACK	1.0	none
16	ACS_PCT_HISPANIC	0.7	sqrt(x)

Figure 2. GGpairs plot of stroke crude prevalence versus select demographic predictor variables



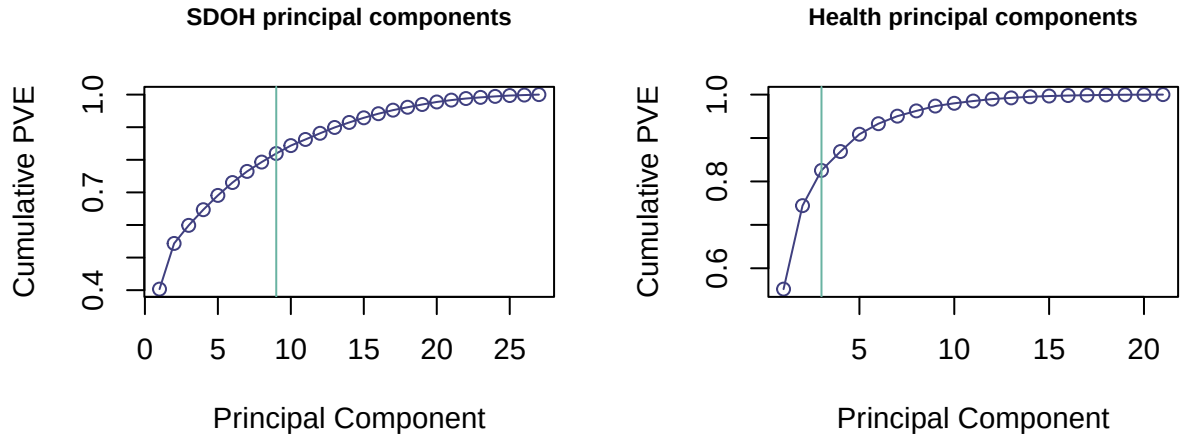
During exploratory analysis, relationships between predictor variables and the response variable were visually examined. A sample plot is shown in Figure 2 (above). A correlation matrix (Figure 3) shows a substantial amount of multicollinearity in the data set. Given the high dimensionality and multicollinearity of the data, the selected modeling methods for this analysis are LASSO regression and principal components analysis (PCA) with linear regression. Both methods are appropriate for handling multicollinearity and reducing dimensionality.

Figure 3. Correlation matrix for stroke crude prevalence and predictor variables



Prior to model fitting, principal components analysis was performed on two key domains using the full data set. The first domain includes variables related to social determinants of health, such as housing, economics, community stability, and deprivation. The second domain includes health related topics, such as access to healthcare, general health status, preventive healthcare, and health risk behaviors. Demographic and environmental variables were not included in the PCA analysis. Using a threshold of 0.8 for the cumulative proportion of variance explained, 9 components were retained for social determinants of health and 3 components were retained for health (Figure 4).

Figure 4. Cumulative proportion of variance explained



Loading values were examined to provide an interpretation for each principal component. Meaningful interpretations were found for most principal components. Social determinant of health PC1 (Figure 5) loaded positively onto percent of households receiving food stamps and percent of households with less than $25,000 in income while loading negatively onto median household income, income per capita, and median home value. This principal component is interpreted as a measure of poverty. Health PC2 (Figure 6) had strong positive loadings for routine health checkups, receiving a cholesterol screening, and taking medicine for high blood pressure. This principal component is interpreted as a measure of preventive health care.

Figure 5. Principal component loadings for SDOH PC1

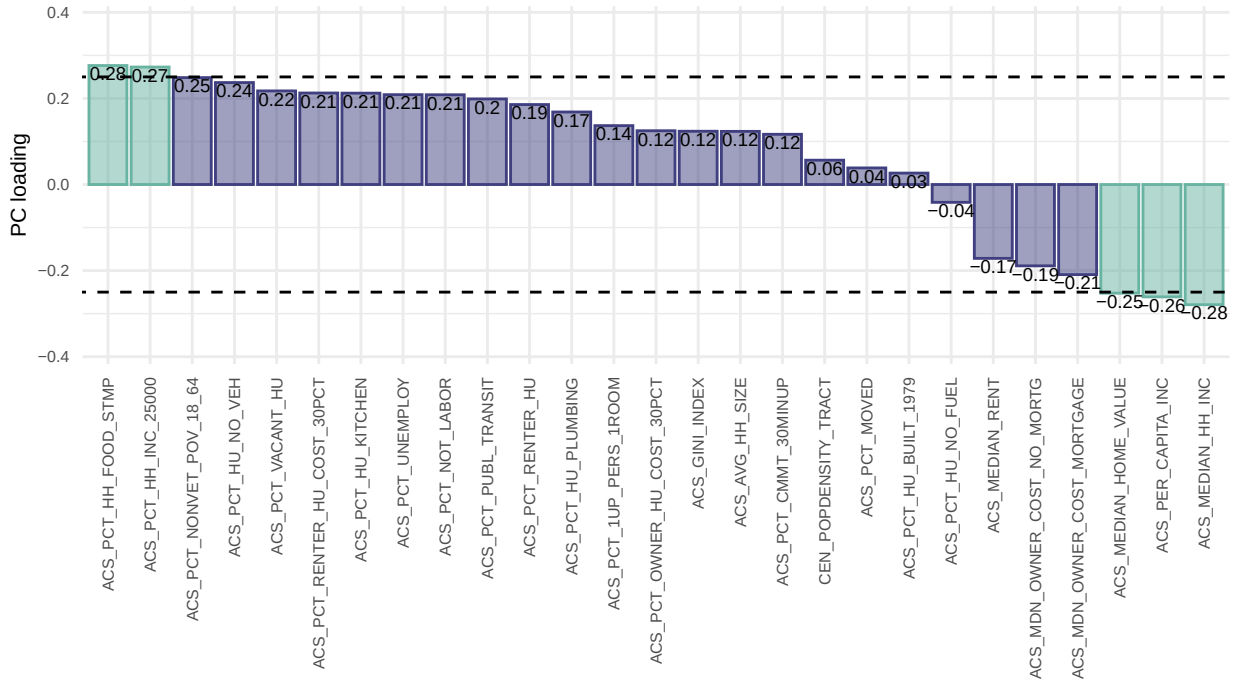
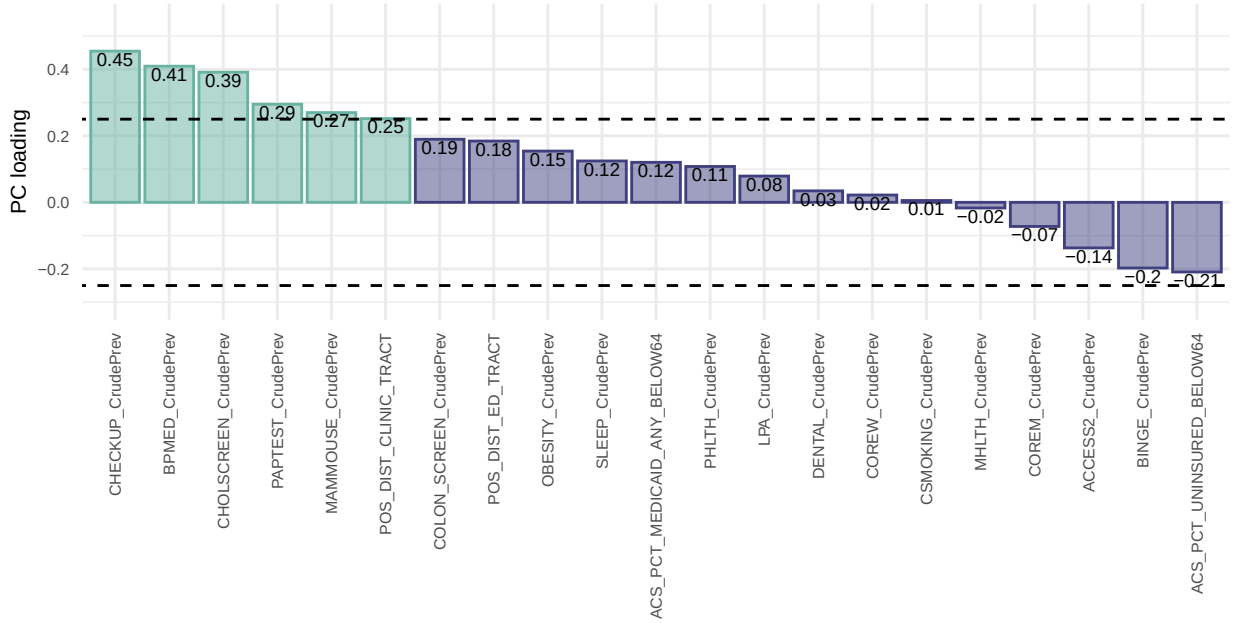


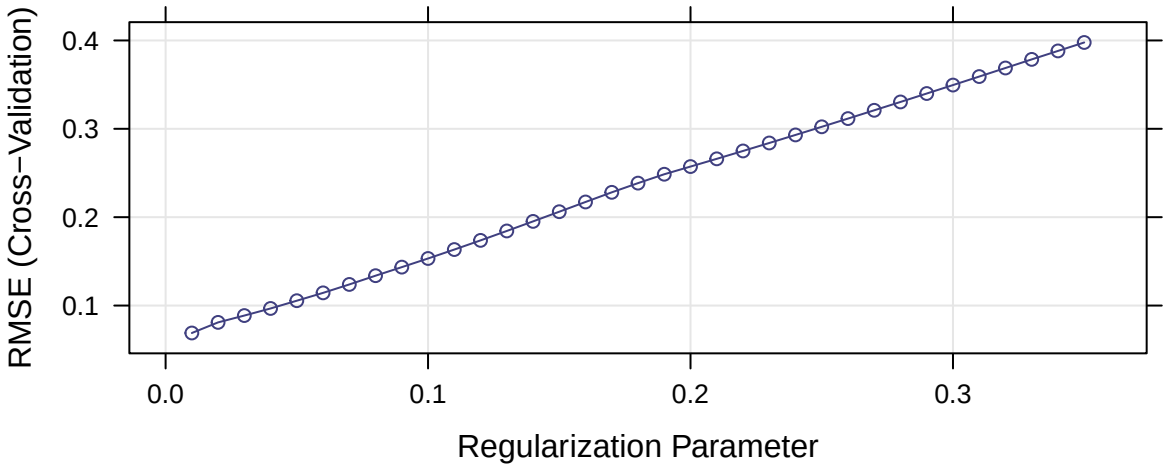
Figure 6. Principal component loadings for Health PC2



After applying PCA, the dimensionality of the data set was reduced from 65 predictors to 29 predictors. PCA also resolved the impediment of multicollinearity, making the final data set suitable for linear regression. The trade off between parsimony and interpretability inherent to PCA may be worthwhile if it produces a superior model for predicting stroke crude prevalence; however, it will likely reduce interpretability of the model and ability to make specific recommendations to reduce stroke incidence.

A single layer of 10-fold cross-validation was used to fit the LASSO regression and PCA linear regression models. A tuning parameter, lambda, was used to fit the LASSO model. The cross-validation RMSE plotted against different values of lambda is displayed below (Figure 7). A lambda value of 0.01 was found to be optimal. Single cross-validation preferred the LASSO regression (RMSE = 0.069) over the PCA linear regression (RMSE = 0.104) as the better model.

Figure 7. LASSO regression: RMSE for different values of lambda



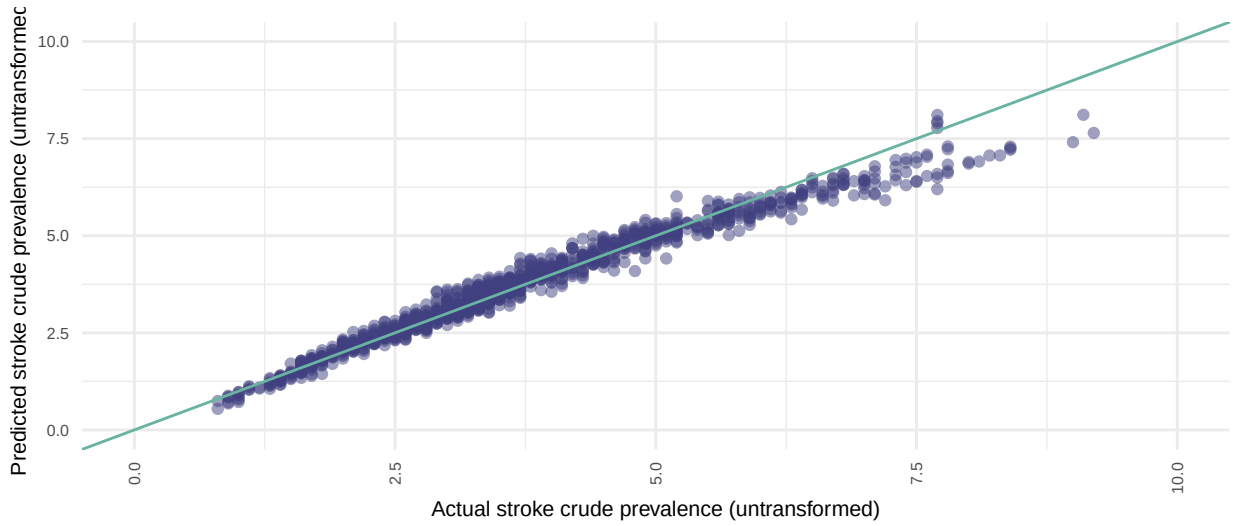
Next, an outer layer of 5-fold cross-validation was conducted to assess the performance of the model selection process. The double cross-validation process consistently chose the LASSO regression as the best model (Table 2). This validates the results of the single cross-validation, which also preferred the LASSO regression over the PCA linear regression.

Table 2: Double cross-validation results table

Best model at outer loop 1 is LASSO linear regression with lambda = 0.01, RMSE = 0.0697, and R^2 = 0.97
Best model at outer loop 2 is LASSO linear regression with lambda = 0.01, RMSE = 0.069, and R^2 = 0.97
Best model at outer loop 3 is LASSO linear regression with lambda = 0.01, RMSE = 0.0706, and R^2 = 0.97
Best model at outer loop 4 is LASSO linear regression with lambda = 0.01, RMSE = 0.069, and R^2 = 0.97
Best model at outer loop 5 is LASSO linear regression with lambda = 0.01, RMSE = 0.0701, and R^2 = 0.97

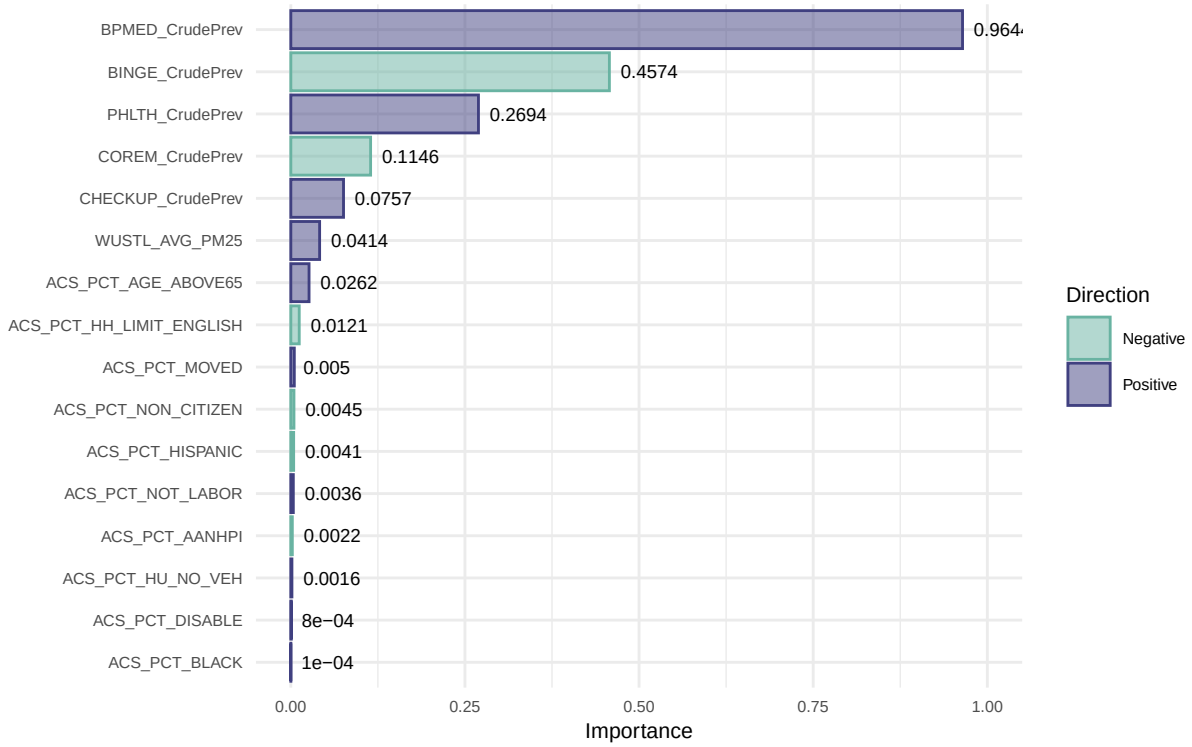
The honest double cross-validation model assessment shows an RMSE value of 0.07. The prediction accuracy of the best model is unusually high. The mean absolute error for the best model indicates the predicted values are, on average, 0.054 units away from the true stroke crude prevalence values. A plot of actual y-values versus predicted y-values (Figure 8) shows a high level of predictive accuracy for census tracts with low to moderate crude prevalence of stroke. Among census tracts with higher rates of stroke, the mean absolute error increases where the model tends to under-estimate stroke crude prevalence. To improve the fit of the model, it is recommended that more census tracts with high stroke crude prevalence rates be included to increase the number of observations training the model. Including additional years of data may certainly assist with this. It is also possible to include the full PLACES data set to increase the overall number of census tracts (see Footnote 1).

Figure 8. Predicted vs actual y-values for LASSO regression



Finally, the best model (LASSO regression) was fit to the full data using the optimized control parameter (lambda=0.01). The LASSO model identified three predictor variables of greatest importance when predicting stroke crude prevalence (Figure 9): crude prevalence of blood pressure medication use, crude prevalence of binge drinking, and percent of adults who report 14 or more days during the past 30 days during which their physical health was not good. Use of blood pressure medication and reporting poor physical health were positively related to crude prevalence of stroke. Binge drinking was negatively associated with rates of stroke at the neighborhood level.

Figure 9. Variable importance for LASSO regression



The predictors of greatest importance are both surprising and unsurprising. Higher prevalence of blood pressure medication usage indicates a higher prevalence of cardiovascular disease in a census tract. Cardiovascular disease is a significant risk factor for stroke. On an individual level, binge drinking is a known risk factor for stroke. However, in this analysis, higher prevalence of binge drinking was associated with lower prevalence of stroke. This may be explained by the lower prevalence of binge drinking in census tracts with incomes below the median. Household income and poverty are strongly associated with many adverse health outcomes, including stroke.