Predicting County-Level Heart Disease Mortality in the United States

Charlotte Abrams, Laura Cosgrove, Alyssa Vanderbeek

7 April 2019

Introduction

Heart disease remains one of the leading causes of death in adults in the US. Understanding risk factors for diseases of this kind is an important task in working towards reducing the number of lives lost. There is obvious benefit in doing this at an individual level, examining personal lifestyle, genetic profile, and family history. But there may be important environmental components that have predictive value in assessing risk for heart disease mortality. Identifying county-level environmental and demographic predictors for heart disease mortality could aid in effective targeting of national campaigns to increase education and reduce risk. We examine economic, health, and demographic data for thousands of counties across the US. This data is synthesized from several sources, including the USDA Economic Research Service, Bureau of Labor Statistics, US Census, Behavioral Risk Factor Surveillance System, the CDC, and others. Our goal in this project is to most effectively predict the county-level heart disease mortality rate per 100,000 persons. We build 6 predictive models (stepwise linear regression, Lasso, Ridge, PCR, GAM, and MARS), and compare them on their predictive capacity quantified by the root mean squared error (RMSE).

Exploratory Data Analysis

The average county-level heart disease mortality per 100,000 residents is 279.37, with standard deviation 58.95. Though approximately normally distributed, the exact density of the outcome distribution differs slightly among levels of categorical variables.

In the category of health, statistics related to lifestyle characteristics (obesity, smoking, diabetes, low birthweight babies, excessive drinking, and physical inactivity) are highly intercorrelated (Figure 1). Less so are more environmental - or "acts of God" - factors: air particulate matter, homicides, motor vehicle crashes, and rates of dentists and doctors (though the last two are highly correlated). For demography, we found categorical variables masquerading as separate predictors, leading to high intercorrelation within those categories: percentages of residents who are a given race accounts for 5 variables; age-related bins (less than 18, greater than 65) account for 2 variables; birth and death rate; and percentages of residents who complete a given level of education account for 4 variables. Given some of the subtleties in these distinctions, we condensed only race into "white" and "non-white". For economics, unsurprisingly, the percent of adults and the percent of children without health insurance are highly correlated, as well as percent civilian labor and unemployment rate (Figure 1).

Between categories, education-related variables are highly correlated with % civilian labor, % uninsured adults, and % physical inactivity. Predictors strongly associated with heart disease mortality are % physical inactivity, % diabetes, % adult obesity, education-related variables, % low birthweight, and overall death rate per 1,000. Many of the variables most strongly-associated with the outcome are associated with one another.

Methods

In our data pre-processing and exploratory analysis, we examined variable distributions (checking for skewness, missing data, etc). We eliminated two variables with over 90% missing data, and used K-nearest neighbors (KNN) to impute remaining missing values. The dataset was split into training and testing subsets at a parsing 2:1 ratio. There was one categorical variable for which individual factor levels had near-zero variance,

but because these levels were part of a larger parameter, we did not remove them. The final dataset includes the outcome (heart disease mortality rate per 100k) and 28 predictors relating to economic, demographic, and health-related county characteristics.

We built predictive models on the training data using 6 methods: stepwise linear regression, Lasso, Ridge, PCR, GAM, and MARS. All model builds were performed using the `caret` package, and optimal model parameters were selected based on minimizing MSE (or, in the case of MARS, and GAM, generalized cross-validation error) with 10-fold cross-validation, repeated five times. The optimal model was selected based on training RMSE, and models were again compared based on test RMSE. All R code used for model fitting is provided in the Appendix.

Results

Predictive Models

Final models and RMSE

All models used the majority of all 28 variables (where categorical variables were split into dummy variables).

Table 1: Predictive RMSE on training and test data

	Train RMSE	Test RMSE
Stepwise	33.85305	32.90754
Ridge	33.67035	33.04168
Lasso	33.66725	32.89532
PCR	33.61412	32.86776
GAM	32.50513	30.67289
MARS	33.76163	30.83814

Above, table 1 presents the RMSE of the predicted heart disease mortality rate for all models on both the training and testing datasets. Figure 2 presents the resampled training RMSE. GAM outperforms all other models in both fields, though MARS has similar performance on predicting outcomes in the test data. All linear models have similar performance.

Coefficient Shrinkage: Lasso and Ridge

For better visualization of coefficient shrinkage among the possible , a `glmnet` model was fit and the lambda value selected by `caret` as minimizing RMSE through cross-validated iterations was plotted for ridge (Figure 3).

Investigation of MARS

The minimum generalized cross-validation error was achieved for a total of 25 features, including a two-way interaction term. Variable importance is assessed by tracking GCV for each predictor and accumulating the reduction in GCV when each predictor's feature is added to the model – the total reduction is used as the measure of variable importance. If a predictor was never used in any MARS basis function, it has an importance value of zero; 18 predictors were used in a MARS basis function. MARS retained interactions between many sets of health-related variables, which one might expect given collinearity (and relatedness) between the within-category variable sets. Interaction terms were also present across categories.

Investigation of GAM

GAM fit 23 regression splines with a total degree of freedom of 78.05. Two splines had degrees of freedom close to 0, and therefore the predictors are effectively not included in the model: percent unemployment (econ___pct_unemployment) and percent adults with highest education of some college (demo___pct_adults_with_some_college). Additionally, four variables in total had p values > 0.05:

percent unemployment, percent adults with highest education of some college, percent adults with bachelors degree or higher (demo___pct_adults_bachelors_or_higher), and total population per dentist (health___pop_per_dentist).

An example regression spline fit for GAM is shown in Figure 5. For percent residents below 18 years of age, change heart disease mortality varies from 4 standard deviations below average percent residents below 18 years of age to 3 standard deviations above, then sharply increases from 3 standard deviations above to 6 standard deviations above average percent residents below 18 years of age. Another example is for percent nonwhite residents, where heart disease mortality increases from a standard deviation lower percent nonwhite residents to an average percent nonwhite residents, but from an average level of percent nonwhite residents to three standard deviations higher, heart disease mortality decreases.

Compared to MARS and PCR, GAM gave more weight to the factor variable metropolitan status and metropolitan adjacency.

Discussion

GAM was our best-performing model. With GAM, we were able to achieve a cross-validated RMSE of median 32.45 and bootstrapped interquartile range of (31.18, 33.34. GAM, and MARS, also performed well on the test data. All models had similar distributions of residuals.

This modeling problem was perhaps not best addressed by fitting lasso and ridge. Neither of the methods shrunk many coefficients in such a way that completely removed them from the model and improved the RMSE. The best visualization of the effect of shrinkage can be seen in the ridge coefficient plot (Figure 3), where it can be seen that the vast majority of coefficients shrunk equally. Lasso's more stringent shrinking power did not improve RMSE for this problem. Had these methods been more appropriate in this context, parameter selection might have been more obvious. As it is, the standard error of the mean cross-validated error was quite wide.

As described, GAM and MARS outperformed the linear models, and if the primary goal of the model is prediction, than we suggest using GAM. An even more flexible model might better capture the underlying truth, because the RMSE was still quite large. However, the linear models were close in performance to these more flexible models, and if simplicity and interpretability are major concerns, one of these may be preferred. For the linear model fit with stepwise selection, for every standard deviation increase of death rate per 1000 residents, heart disease mortality per 100k increases on average by 19.3 (95% CI 16.9 – 21.6). Surprisingly, percent residents aged 65 years and older has a decreasing effect on heart disease mortality: for every standard deviation increase in percent residents aged 65 years and older, heart disease mortality per 100k decreases by 19.8 on average (95% CI -23.0 – -16.6). Percent low birthweight increases heart disease mortality on average (1 s.d. change increases by 7.4, 95% CI 5.3 – 9.4), and percent physical inactivity increases heart disease mortality on average (1 s.d. change increases by 9.0, 95% CI 6.5 – 11.6).

Variable importance was relatively consistent through each model, as measured by loss-perturbed RMSE in the `caret` cross-validation and shown in Figure 4, generated by the package `DALEX`: for instance, overall death rate in a county consistently had a large, positive effect on the rate of heart disease mortality. However, compared to other models, GAM gave much more weight to the value of the geography-related factor variables metropolitan status (metro) and adjacency (metro_adjacency). Stepwise regression retained urban influence, and not other geography-related variables, and as we noted in exploratory data analysis, urban influence serves somewhat as a proxy for an interaction term between population and metropolitan status. Although the importance of particular predictors in a model with finite CV error-based parameter selection is subject to some randomness (or non-uniqueness), and therefore the meaning of importance should not be over-interpreted, it is possible that including several proxies for the same predictor type over-complicated predictor selection, particularly in the lasso, ridge, and PCR cases. A different approach to try would be to begin with fewer factor predictors and note whether lasso, ridge, and PCR would have more effective shrinkage.

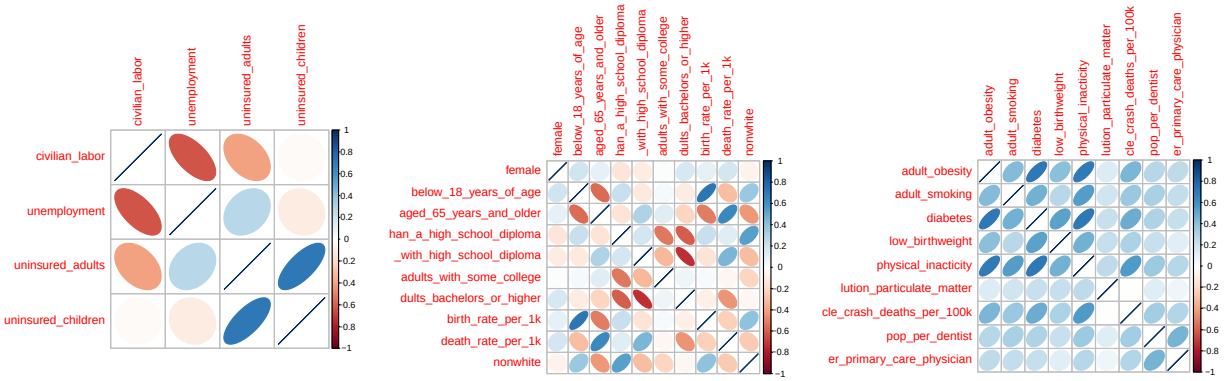

Figure 1: Correlation within predictor categories

Tables and Figures

```
## integer(0)
```

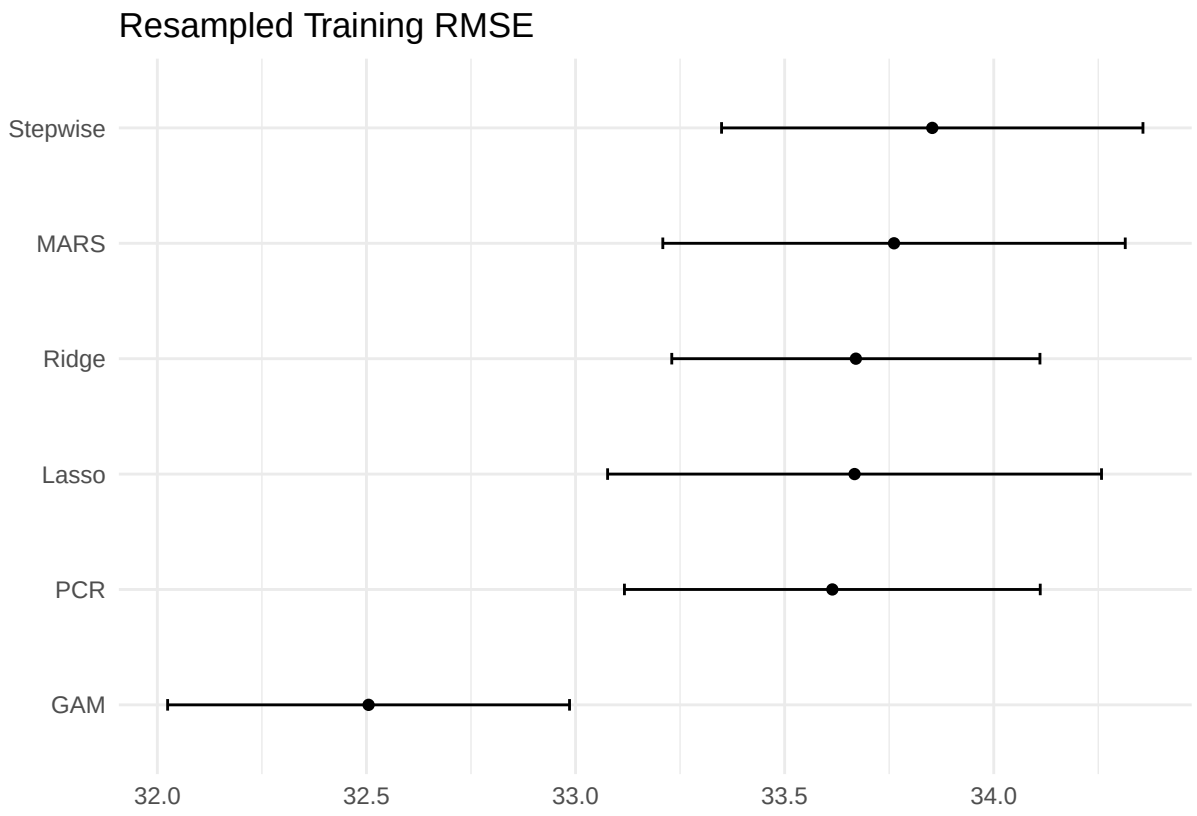

Figure 2: GAM performance exceeded other model performance

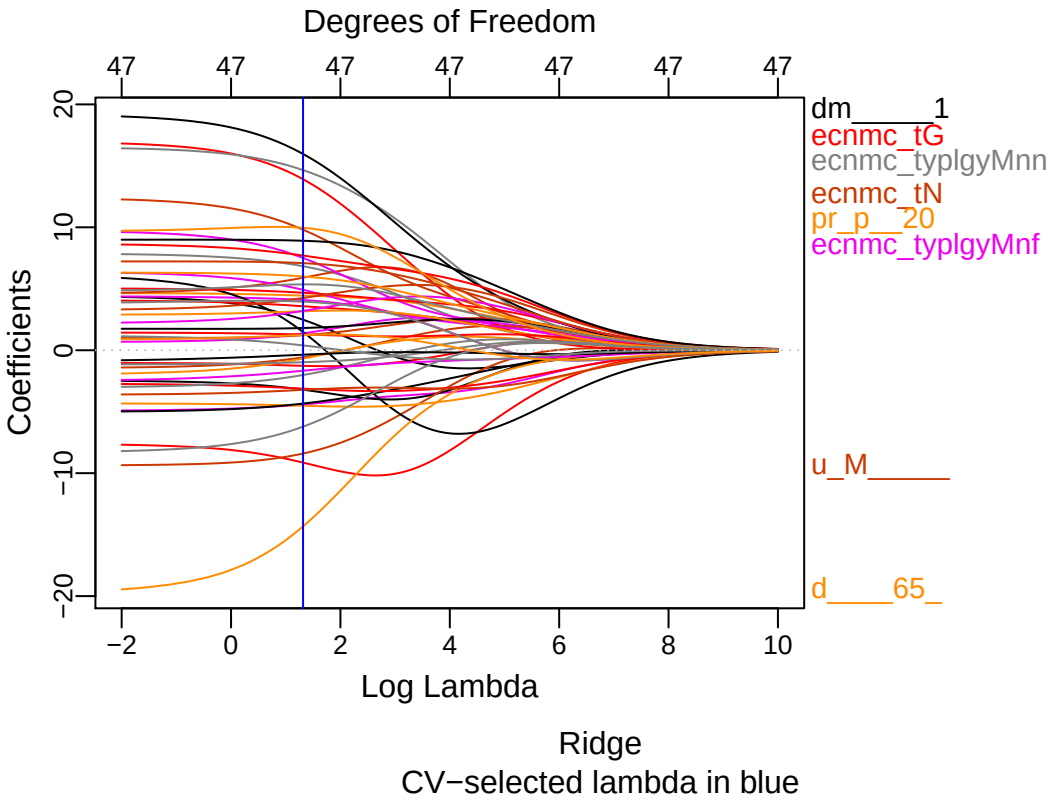

Figure 3: Shrinkage performance of lasso and ridge was poor: CV selected small shrinkage (small lambda)

Figure 4: GAM performance relies heavily on the weight of metropolitan status

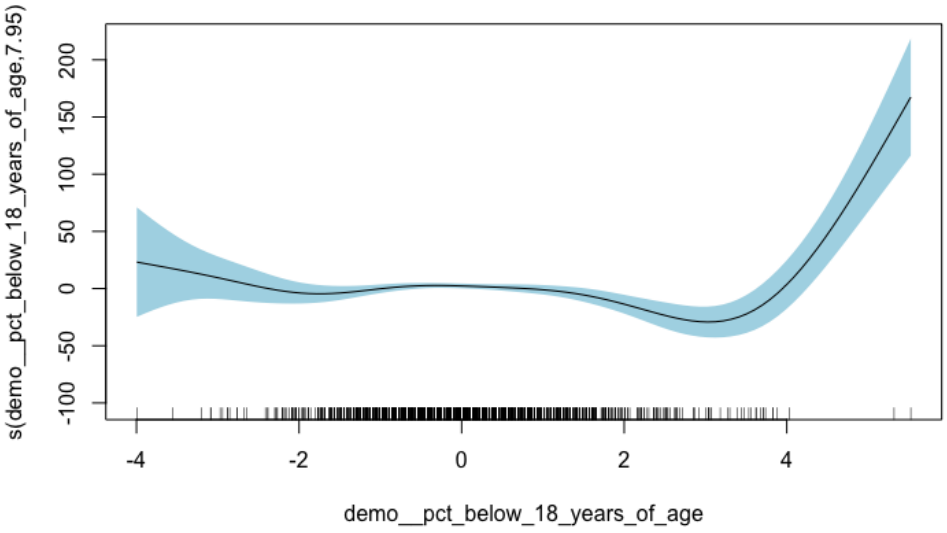

Figure 5: GAM smoothing splines fit