An Analysis of Health in U.S. Citizens

Graham Patton, Ryan Brunner, Troy Hall

The University of North Carolina at Chapel Hill

Economics Department

ECON 573 Research Paper

Professor: Andrii Babii

# Introduction

Researchers across a diversity of disciplines have claimed a strong correlation between income inequality and measures of health outcomes on an individual basis. Richard Wilkinson stated that “It is now clear that the scale of income differences in a society is one of the most powerful determinants of health” [1]. Each year, the U.S. Bureau of Labor Statistics and the U.S. Census Bureau jointly sponsor the Current Population Survey (CPS), which serves as a primary source of information on labor and population statistics for the US government and a litany of social science researchers. This paper intends to research the previously mentioned statement that income serves as a primary driver in health outcomes. In 2002, researchers Jennifer Mellor and Jeffrey Milyo analyzed CPS data, finding no “consistent association between income inequality and individual health status” [2]. This research aims to explore the correlation between the two variables: our null hypothesis being that there is no statistically significant association between income and health status. In 2002, Mellor and Milyo found that when individual characteristics were controlled for, no results were statistically significant enough to allow them to reject the null hypothesis that income did not affect health. Our research aimed to investigate the validity of their results by introducing several factors that we deemed necessary to the model.

# Literature Review

Our focal paper is a 2002 piece by Mellor and Milyo, in which they built an OLS regression predicting the incidence of poor or fair health (lowest two levels of fair and poor versus the upper three on a surveyed five-point scale of excellent, very good, and good), using race, Hispanic ethnicity, sex, marriage, health insurance coverage, central city status, metropolitan area status, education, age and income. Their OLS models determined that income was not a significant variable in health outcomes. Using CPS data, our research aims to use similar variables with updated more sophisticated models to explore their same hypothesis and evaluate what covariates are most significant in predicting the health of an individual. Using the CPS survey, our included models will be linear, LASSO, ridge and elastic net regression in addition to regression trees, random forests, boosting, support vector margins and K-nearest neighbors.

# Data

The cross-sectional data used to build our models comes from the March 2019 Annual Social and Economic Supplement to the Current Population Survey. CPS data is designed to be representative of the population in its sampling scheme, and has the advantages of a high level of quality with a large number of observations thanks to the Census Bureau’s financial and political firepower. The Census Bureau is incentivized to keep high quality records because many key decision makers in policy and in social science rely on it to come to conclusions about the state of the US. One major limitation of the data set lies in its representative nature. Observations for minorities and fringe cases remain low and have the ability to skew results, especially with the growing number of US citizens who identify with a minority group. For this reason, we subset the data to allow for greater homogeneity across major factors like race, limiting the racial factors to only four categories. Furthermore, the factor variables for industries, marriage status, nativity, and education were all multi-leveled factor variables with what we deemed to be likely homogenous effects to the target variable. When possible, these were reconciled to binary variables to reflect these homogenous effects. Variable descriptions can be found in the Appendix.

# Model Specification

It is important to acknowledge some assumptions and transformations that we have made to the data to ensure that we get the most accurate results possible and are able to draw meaningful conclusions from our analysis. Since the surveys were filled out by heads of households, we subset to include only individuals who are over the age of 18 as the health status of children will likely depend heavily on whatever their parent’s overall physical and mental health looks like. Another necessary assumption is the accuracy of the survey. A five-level subjective scale of overall health will not capture all the nuances of chronic illness, the effect of diet and exercise on health, or all the other factors that contribute to whether a person considers themselves “healthy” or not. We conjecture that on average, people will have similar ideas on what health looks like and we can rely on a subjective health score so long as our sample size remains large to take care of outliers.

Unless otherwise stated, all models will have the variables included in the Appendix table of Variable Descriptions. We will denote if the variable HEALTH takes the form of a continuous variable or a factor depending on the model. Theoretically, our model looks at health at a snapshot in time. We have no way of knowing how this may change health over time - we can only tell which variables are correlated with subjective levels of health in 2019. The COVID - 19 pandemic may have also drastically changed how people perceive their health, so we elected to use data from the last period before the pandemic began to get a sense of determinants of health without the added turbulence of COVID.

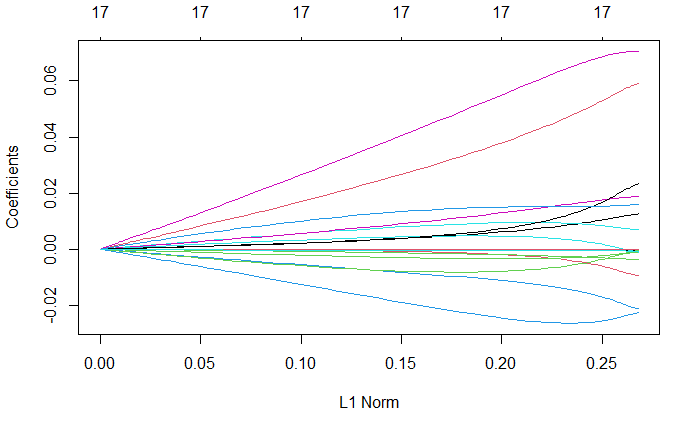
# OLS

To determine individual effects of race and levels of education, those variables were factored in order to separate their leveled effects. Seeing as Mellor and Milyo used regression, it seemed most pertinent to start with their own analytic methodology. With health being the target variable, we regressed it on all other factors to determine significant variables and determine the greatest predictors of an individual’s health. OLS, being one of the simplest analytic methodologies, is one of the most interpretable ones. Like Mellor and Milyo’s analysis, income did not serve as a large determinant of health in an individual. It was one of the most significant variables at a near zero p-value, but it did not contain the largest effect among variables of significance. Higher levels of education and indication that the individual was in the labor force were both significant and predicted to have greater effects on health outcomes as compared to health, but not by large amounts. As for insignificant variables, OLS regressions predicted that being Black, being a female, being in a metro area, moving in the past year, losing a job in the last year and having health care coverage were not statistically significant in predictions of health outcomes. The statistically insignificant indicator variable of healthcare coverage comes as a huge surprise and may be a result of the mostly subjective health variable we use on the LHS. Income is not the largest predictor of health outcomes in our model, which aligns with what Mellor and Milyo found. However, it is a highly significant variable that has an undeniable effect on health, with a one unit increase in logged income producing a 0.05 increase in health. This is still a massive effect, especially considering health is bounded between 0 and 1. The R-squared value was 0.202 which indicates that around 80% of variation in the data cannot be explained by the model, however the significance of the high F statistic indicates that the model does hold predictive power. The adjusted R-squared was only 0.004 below the multiple R-squared. The Mean Squared Error of this model was 0.05804, which seems extremely small but makes sense when considering our dependent variable is bounded between zero and one. Results for individual coefficients are reported below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.0788211471937765 | 0.0149358123806108 | 5.27732574467123 | 1.31309476104113e-07 |
| factor(RACE)Black | -0.00620490214275673 | 0.00394359269302181 | -1.57341354083963 | 0.11562572698727 |
| factor(RACE)Other | -0.0102568194617148 | 0.00499298281569597 | -2.05424689816103 | 0.0399539600558749 |
| factor(RACE)White | 0.0239368940822344 | 0.00352229130411315 | 6.79583033188711 | 1.08189984885823e-11 |
| HISPAN | -0.00959399643184835 | 0.00242014703986029 | -3.96422046835725 | 7.36787438325346e-05 |
| FEMALE | -0.000409117806001658 | 0.00145040736305009 | -0.282070966008692 | 0.77788958131281 |
| MARST | -0.022833955137986 | 0.00161583542622921 | -14.1313618746882 | 2.6464616555964e-45 |
| METRO | 0.002044219205276 | 0.00193203076716784 | 1.05806762501646 | 0.290026740454147 |
| Education2 | 0.0250188483368805 | 0.00258232787470169 | 9.68848633900556 | 3.43868729500034e-22 |
| Education3 | 0.0480353876775204 | 0.00283228715880162 | 16.9599284903883 | 1.93258805709138e-64 |
| Education4 | 0.0485746404088743 | 0.00320222770177268 | 15.1690151146917 | 6.33799627414081e-52 |
| Education5 | 0.079316600433303 | 0.00287317576182183 | 27.6058991890596 | 3.15947806186465e-167 |
| Education6 | 0.0963743336577019 | 0.00323870384071317 | 29.7570689997021 | 7.15139523086349e-194 |
| HHINCOME | 0.0588534658601374 | 0.00115120771517245 | 51.1232378696501 | 0 |
| AGE | -0.00380191540240004 | 4.89237347810181e-05 | -77.711062318062 | 0 |
| VETSTAT | -0.021511564313172 | 0.00269643624546605 | -7.97777598092402 | 1.50293800541119e-15 |
| FAMSIZE | -0.00134999641893675 | 0.000521995952962866 | -2.58622008709863 | 0.00970465993388215 |
| LABFORCE | 0.0711442510701265 | 0.00184154256201934 | 38.632965937053 | 0 |
| bornabroad | 0.0175693105069779 | 0.00238466243060439 | 7.36763001819299 | 1.7479512673642e-13 |
| UHRSWORKT | 7.41379021032553e-06 | 1.84892769473965e-06 | 4.00977833336499 | 6.08124617985935e-05 |
| indmove | 9.77358118536873e-05 | 0.00181749262446678 | 0.0537750803155865 | 0.95711445322804 |
| lostjob | 0.0128530020132377 | 0.0171530450050481 | 0.749313139996724 | 0.453669956992857 |
| ANYCOVNW | 0.00057043135278613 | 0.00245579324868627 | 0.232279876610656 | 0.816321036852125 |

# Ridge

Ridge regression is an analytic technique for multiple regression data. It is designed for highly correlated data, where all coefficients are shrunk towards zero to trade off varying levels of bias for a lower variance. In OLS, when multicollinearity occurs, variances can be so large that predictive power decreases, despite the unbiased nature of the coefficients. To increase predictive power, ridge regression allows for bias to decrease variance and maximize predictive power. Lambda.min gives the value of lambda that produces the lowest cross validated error while lambda.1se gives the highest regularized model within one standard error of the minimum value of lambda. The lambda.min value was had a MSE of 0.008390432 and the 1se value MSE was 0.05393431. We decided to move forward with the minimum value as our objective is to minimize our test error. The R-squared for the model was 0.1975985 which indicated only slightly better explanation of variation in the data over OLS. At the lowest MSE, the variables estimated to have the greatest predictive power for health outcomes were income and labor force participation. From the graph below, very few coefficients seem to be heavily penalized and there are only a couple of possible standout predictors that may have great predictive power.



# LASSO

Lasso, like ridge, uses shrinkage to bring coefficients towards zero. However, unlike ridge, it produces sparse solutions (model selection) by actually bringing some coefficients to zero, and therefore does a better job of highlighting the most important variables. “Lasso” stands for “least absolute shrinkage and selection operator.” With alpha set to 1, the ideal lambda was found to minimize MSE to 0.0002177337. the R-squared at this best lambda was 0.1976951, which is marginally better than that of the ridge regression. The model deemed that indication of a move was unnecessary for the model and brought its coefficient to zero in its sparse solution. Lasso, like ridge, deemed that income and labor force participation were the most important factors in prediction of health outcomes. It is interesting to note that after running a principal component analysis, it was not found that any single component explained less than two percent of the variation in the data. The most that any one component explained was around 15%. This might explain why Lasso did not bring more coefficients to zero. Each variable, while not explaining a massive amount of variation, likely explains enough to prevent the lasso from shrinking it to zero. Graphs for both how predictors fared under the lasso and optimal models at different levels of log-lambda are provided in the following space.

# Elastic Net

A generalized model between Lasso and Ridge, Elastic Net regression’s advantage lies in an elastic net penalty that allows it regularization like Ridge and create sparse solutions like Lasso. After training the model, the elastic net regression determined the best alpha value to be 0.1968718 and the best lambda to be 0.009019538. The RMSE for the model was 0.2414369. The R-squared for this model was 0.2009366 (MSE = 0.0403755). As for variable selection, seeing as elastic net automatically factored certain variables, it created technically sparser solutions than Lasso. This regression found that gender, Hispanic ethnicity, many of the state effects, family size, indication of a move, indication of a lost job, and indication of healthcare coverage were all insignificant factors to the model. This model also has improved prediction power over the Ridge regression and Lasso models. Elastic nets with different regularization parameters/percentages between Ridge and LASSO are shown graphically below.

# 

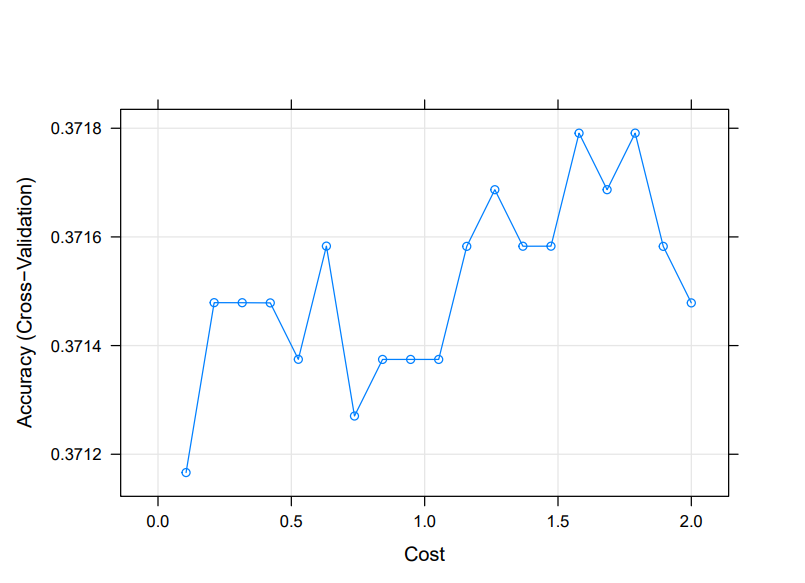
# K - Nearest Neighbors

K - nearest neighbors, also known as KNN, is an incredibly flexible machine learning procedure that involves exactly what the name implies. Each point is classified based on its K - nearest neighbors, with the number of nearest neighbors used for classification an arbitrary number determined by the analyst. “Nearest” here is determined by the Euclidean distance between the point being classified and the points around it based on the variables of each observation. KNN has very few diagnostic plots that can be used to check the importance of each variable or how good the model is at classifying observations. We can, however use the KNN model to test the trained model on a held-out sample of data. We report a table of results below for levels of K from 1 to 10. Test error gets consistently better as values of K increase as more information gets considered for use in the model. Better is not quite good, however, as the test percentage correct increases by about 0.002 to 0.003 for each increase in K that starts at a low end of 0.309 at K = 2. Test accuracy throughout this project is fairly low and likely reflects the difficulty in predicting health outcomes for any given person in the US solely based off of demographic and economic variables.

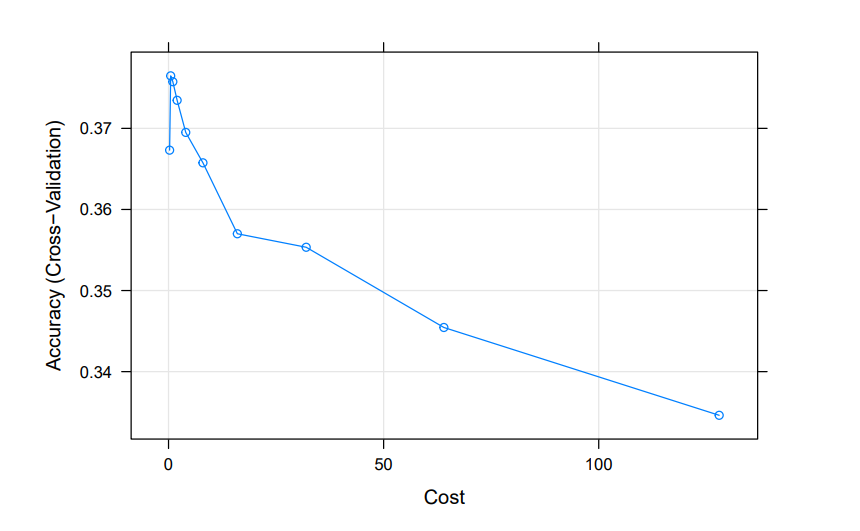
|  |  |
| --- | --- |
| K | Test % correct |
| 1 | 0.3171661 |
| 2 | 0.3097952 |
| 3 | 0.3221577 |
| 4 | 0.3221076 |
| 5 | 0.3337049 |
| 6 | 0.3354021 |
| 7 | 0.3391957 |
| 8 | 0.3428895 |
| 9 | 0.3454019 |
| 10 | 0.3479643 |
| 100 | 0.3707925 |

# Support Vector Machines

Support vector machines create hyperplanes through data subspace in order to separate groups of data. The first step was to factor health outcomes to separate the outcomes and their predictive vectors. The optimal kappa value was 0.1085 with an optimal C-value of 1.789474. The C-value is the parameter in which the margin is set, separating classes within the data. By comparing the trained model to testing values, the model achieved an accuracy of 0.3757298 with a near 0 P-value of 3.971e-08. This was done with 10-fold cross validation. The 95% confidence interval for accuracy was [0.3563, 0.3955].

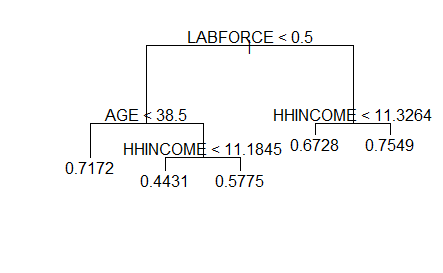


Upon changing the kernel to radial for support vector testing, the optimal C-value was 0.5, with a sigma value of 0.009569431 and kappa of 0.1006. It an accuracy of 0.3745 with a confidence interval of [0.3551, 0.3942]. This was also found by 10-fold cross validation and had a P-value of 2.866e-06. We were unfortunately unable to run polynomial kernel support vector machines due to computational and time constraints. It is also worthwhile to mention that both of these methods, while extremely computationally expensive and requiring us to run them on small subsets of our overall sample, are only marginally better than KNN at predicting the class of health that a given person will fall into.



# Trees

We divided our data set in such a way that we had 25% of our data in the training subset and 75% in the test subset for the tree and forest-based prediction methods. 25% ended up being the largest percentage of the data we were able to use that would allow the models to run in a reasonable amount of time. We initially ran the tree-based method with no subset selection in order to create a model using all relevant variables. The method produced a tree with 5 terminal nodes, with an initial split checking if the variable LABFORCE <.5. LABFORCE is a binary variable that represents whether an individual is a participant in the labor force, with 1 being a participant and 0 not participating. After the branch at the variable LABFORCE, the variables HHINCOME and AGE are also used the further divide the tree into the 5 nodes. Each of these nodes predict a different value for HEALTH. When attempting to trim the tree using cross validation, we found that out tree with 5 terminal nodes produced the lowest MSE. However, as displayed by the graph the regression trees proved to be a fairly unreliable method of predicting health using the variables to which we had access. The root MSE of our best regression tree was .248, meaning that our predicted health value was an average of .248 away from the true value. For reference, had we just assumed that everyone's health was .65 (the mean value of the health column in our data), then our root MSE would have been .270 meaning our model is roughly 8% more accurate than the mean at predicting an individual's health score. This further validates our conclusion that there is little direct correlation between most of our listed dependent variables and Health. The graph of our tree is displayed below.



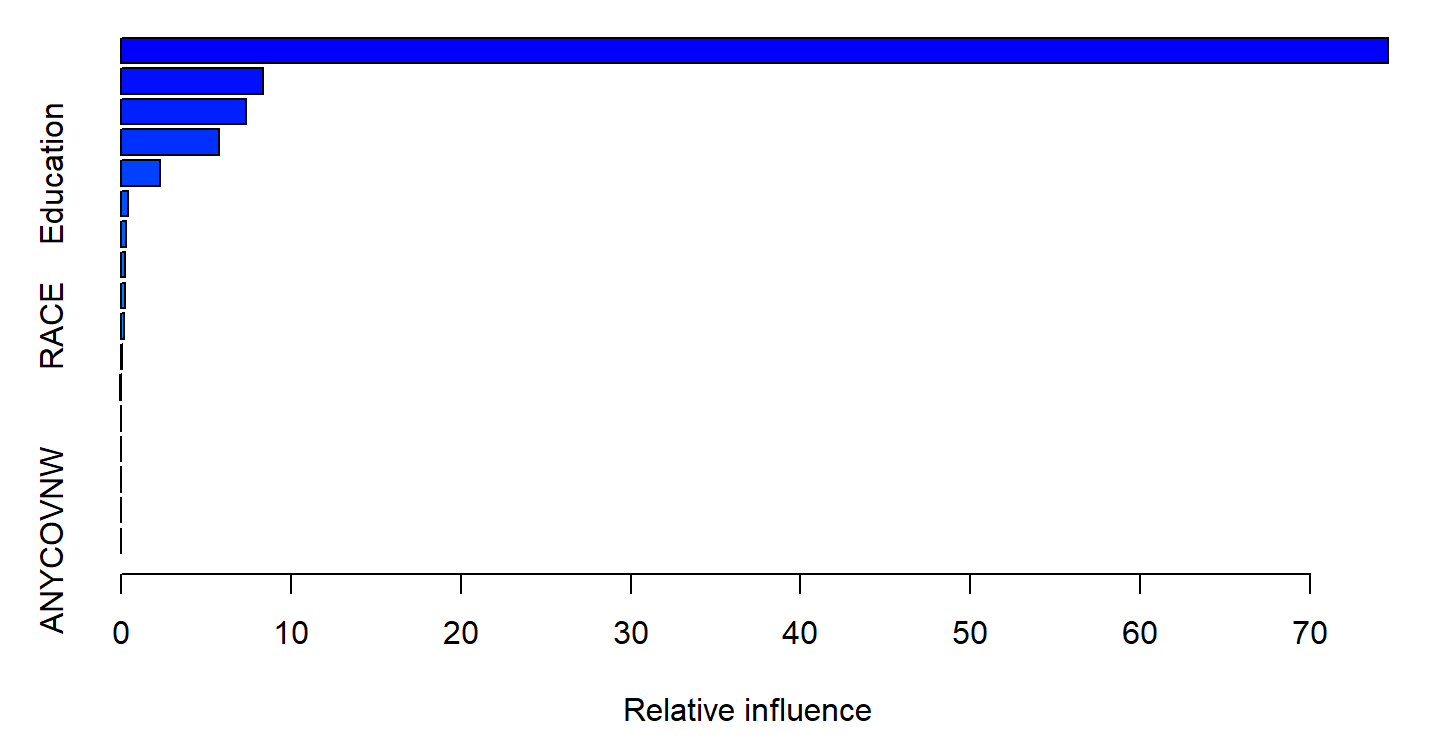
# Random Forests

Using the same 25:75 split we used for the regression trees, we ran a random forest regression on the data set. Given that other models have shown many of our other variables are insignificant, we ran the random forest function with the mtry parameter set to the conventional value for a regression with 18 variables, which is 6. Based on the results of our regression, it appears that the 2 most significant variables by far were AGE and HHINCOME. This is relatively consistent with our other tree-based methods, although the variable LABFORCE appears less significant than in the regression tree where it was the initial break. The root MSE for our random forest was .238 which proved to be a little better than our regression tree, but still supports our initial prediction that none of our variables are terribly significant in predicting one’s health, especially when trying to find cutoffs at these variables that reflect actual levels of health. This isn’t to say that these variables don’t matter; rather, the relationship that each of them has individually to health is fairly weak. When compounded, these variables likely have some sort of effect on predicting health as reflected by our models performing better than just assuming the mean. Graphs displaying the importance of each variable in this process are below.



# Boosting

Boosting is an extension of the tree-based method. Boosting allows us to grow decision trees slowly over time. It attempts to avoid overfitting by fitting new trees to the residuals of the previous model. Using 5000 trees and a tuning parameter of .001 we predicted health using all of our possible predictors. In our boosted trees STATEFIP proved to be the most important variable with AGE and HHINCOME being the next 2. STATEFIP was not a prevalent variable in the model before so it was interesting to see it weighing so heavily in the boosted model. This may be indicative of the wealth/power of a state having an outsized influence on the health of its population. However, with a RMSE of .296, the boosted model is worse at predicting individual health than the mean of the health column is. Thus, the boosted model strongly supports our inference that none of our variables are great predictors for health.



# Conclusion

In the end, we found that not many of our variables were good at accurately predicting the value of a given person’s health in levels from 1 to 5. However, given the fact that we often utilized categories or continuous dependent variables with these methods rather than binary ones, we find these results to be a modest success. We were primarily concerned with finding variables that had a strong relationship with health rather than accurately predicting the health of a given adult, but we didn’t necessarily find the proverbial “silver bullet” that would allow us to definitively say that a variable leads to greater or lower levels of health. We cannot even rule out reverse causality here; healthier people may be able to work more or have fewer time constraints that hold them back from maximizing their income.

This isn’t to say that we had nothing to take way. Even if our predictions accuracies were low, some variables emerged that had consistently larger effects on health. Age had a consistent negative relationship with health in OLS and this was reflected as its status as a key predictor in many of the other models. Household income also had a highly significant positive relationship with health in the OLS model that was reflected throughout the other models and into the trees. Even though this statement may not be groundbreaking, we can say with a fair level of certainty that if one is richer, they are likely to be healthier and if one is older, they are likely to be less healthy. Stating that income does not have a significant effect on health also is contrary to almost all intuition that states people with more income can afford better healthcare, better food, and generally can use that income on health-producing goods and services. Our project reflects a complicated reality in which health is based on a variety of factors; most of which cannot be captured in a survey.

# 

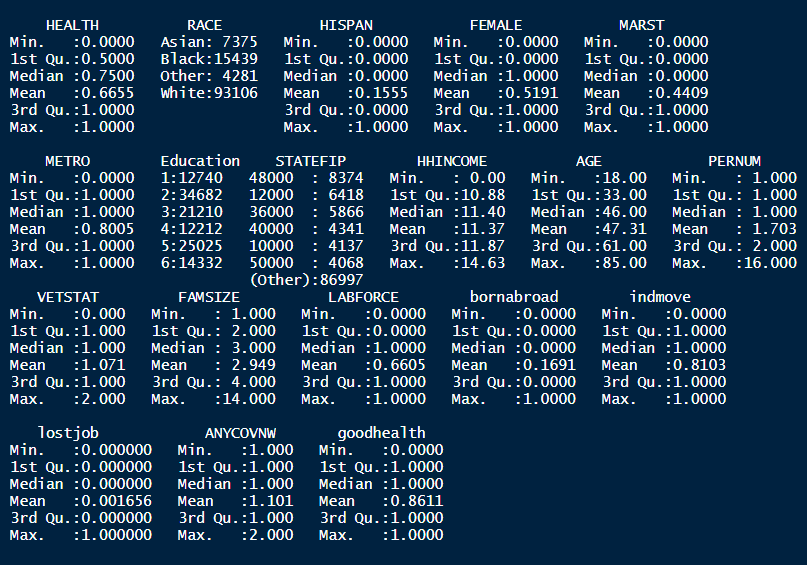
# Bibliography

West, Patrick. (2008). Unhealthy Societies: the Afflictions of Inequality (Book).. Sociology of Health & Illness. 19. 668 - 670. 10.1111/1467-9566.ep10938939.

Mellor, J. M., & Milyo, J. D. (2001). Income Inequality and Health. *Journal of Policy Analysis and Management*, *20*

# Appendix

Summary Stats



Variable Descriptions

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Description | Type |
| MARST | Marital Status, 1 for married and 0 for unmarried | Binary |
| FEMALE | 1 if female, 0 if male | Binary |
| VETSTAT | 1 if veteran of armed forces, 0 if not | Binary |
| LABFORCE | 1 if in labor force, 0 if not | Binary |
| NATIVITY | 1 if born in the US, 0 if not | Binary |
| METROAREA | 1 if individual lives within the metropolitan area of a city listed by the Census as having a metro area, 0 if they do not live in a metropolitan area | Binary |
| Indmove | 1 if individual has changed industries in the last year, 0 if same industry as last year | Binary |
| Lostjob | 1 if individual has gone from employed to unemployed in the past year, 0 if not | Binary |
| ANYCOVNW | 1 if individual had any health coverage over the past year, 0 if not | Binary |
| Race | Race of individual identified in survey. Levels include White, Black, Asian, and Other. | Factor |
| Education | Education level of individual identified in survey. Levels include below HS, high school diploma, some college, Associates, Bachelors, and graduate level training | Factor |
| HHINCOME | Natural logged adjusted household income to eliminate all negative values. Regular household income with the absolute value of the lowest negative income plus one added, then the natural log of the result. | Continuous |
| AGE | Age in years | Continuous Integer |
| STATEFIPS | State level code to control for state - level effects when possible | Factor |
| FAMSIZE | Number of family individuals who live with the survey respondent | Continuous Integer |
| UHRSWORKT | Sum of hours worked by an individual at all jobs | Continuous |
| HEALTH | Variable that ranges from 1 to 0 in increments of 0.25 for each level of health, 1 being the outcome for the healthiest and 0 being the least. Treated as continuous in most models but also used as a factor variable where necessary. | Continuous/  Factor |