Community Socioeconomic Factors Affect on Individual Healthcare Expenditures

Davis Buchanan and Luca Curran

December 19, 2021, 11:59pm

Contents

The code to reproduce this report is available *on Github.*

1 Executive Summary

Problem. Increasingly, the impact of external socioeconomic factors on individual opportunity and livelihood is being examined. Holistic college admissions processes aim to contextualize each applicant's success in order to provide insight into the relative accomplishments of each applicant. Their belief is that success ought to be measured in context with ones socioeconomic circumstances given opportunity access is both meaningful and unequal. Community factors affecting individual healthcare expenses also influence individual economic opportunities.

Accurately predicting future healthcare costs among the medicare-fee-for-service group, those sixty-five or older or suffering from disability and the group with the largest total healthcare-related expenditures annually, will be critical to efficiently target government aid resources. Hence, for our final project we decided to examine various communal socioeconomic factors that affect healthcare cost per capita among this high healthcare-expense group. Analyzing data from the 2020 general presidential election, county-wide health and socioeconomic statistics, and per capita healthcare-related expenditures, we examined which communal socioeconomic and political features best predict greater healthcare spending per capita.

Data. Our dataset pulled from four sources: the United States Centers for Medicare and Medicaid Services (CMS), the Massachusetts Institute of Technology (MIT) Election Data and Science Lab, the University of Wisconsin (UW) Population Health Institute County Health Rankings & Roadmaps, and US Census Bureau (Census) data. Our primary response variable, `per capita health costs` came from the CMS dataset. `Per capita health cost` is defined as the total individual cost of all healthcare-related expenses (out-of-pocket medical services, insurance premiums, personal healthcare goods and services) for the medicare fee-for-service population in USD. This data is from 2019: the most recent data available. The explanatory variables, ranging from `party` (political affiliation of the winning 2020 general presidential candidate in that county) to `household has broadband` (percent of households that have broadband internet subscription), are pulled from the MIT, UW, and Census datatsets. These datasets are primarily from 2020 and 2019. Best attempts were made to use the most up-to-date data where possible.

Analysis. Before exploring our data or running any analyses, we split our data into a training dataset and a test dataset. The test dataset was reserved for assessing and comparing model performance. Following this step, we explored our data, looking at the overall distribution of `per capita health cost,` feature correlation, county presidential election support affect on our response, and the five counties with the highest and lowest `per capita health cost`. From there, we built five predictive models: ordinary least squares, ridge regression, LASSO regression, random forest, and boosting. Boosting had the lowest test RMSE of the tree-based methods, however, OLS had the lowest test RMSE overall.

Conclusions. Variables including the rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees (`preventable_hospitalization`), the percentage of workers who commute alone to work more thirty minutes (`long_commute_perc`), and the percentage of households that spend 50% or more of their household income on housing (`severe_ownership_cost`) were the most robust predictors of high `per capita health cost` across counties. Hopefully, this knowledge can be used to more efficiently target government spending on medicare, inform individual's healthcare plan purchases, and optimize the planning of healthcare insurance businesses.

2 Introduction

Background. Given the necessity of healthcare, its high cost, and varying economic ideologies, policies directed toward reducing the financial burden of healthcare-related expenses remain highly disputed here in the United States. Both Democrat and Republican voters viewed healthcare as their top priority heading into the 2020 presidential election.[1] This is no surprise given the cost of healthcare in the United States is substantially higher than in other countries.[2] Within US borders, however, the cost of healthcare also varies.

[1] Adam Cancryn, 'Politico-Harvard Poll: Health care costs are top priority heading into elections,' *Politico*, 2020. https://www.politico.com/news/2020/02/19/poll-health-care-election-115866

[2] Kamal & Ramirez, Cox, 'How does health spending in the U.S. compare to other countries?' *Peterson KFF*, 2020. https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/

Past studies examining the social-determinants of health show that variation in healthcare costs can be attributed to a host of socioeconomic variables. The COVID-19 pandemic has further exposed inequalities in healthcare access. Key stakeholders in efforts to manage healthcare costs include health insurers, employers, patients, and government agencies.[3] For health insurers, accurate cost forecasts can help with general business planning in addition to prioritizing the allocation of scarce care management resources. The same can be said for government agencies. Moreover, for patients, knowing in advance their likely expenditures for the next year could potentially allow them to choose insurance plans with appropriate deductibles and premiums, ultimately removing some of the financial burden. Accurately predicting future healthcare costs among the medicare-fee-for-service group, those 65+ or suffering from disability and the group with the largest total healthcare-related expenditures annually, will be critical to efficiently target government aid resources, optimize business efforts, and save patients money.

Analysis goals. Given the capacity for a variety of factors to influence individual healthcare spending across counties, we sought to investigate how per capita healthcare costs are affected by various socioeconomic factors. Specifically, we were interested in which variables of interest (e.g. flu vaccine rates, political support, median income, etc.) best predict individual healthcare costs (i.e. per capita health cost).

Significance. Our hope is that this analysis will add to the tremendous amount of research examining the social-determinants of health and vast inequalities in access to affordable healthcare exposed by the current pandemic. Accurately predicting the cost of healthcare will result in improved allocation of government and non-profit resources designated to assist the medicare-fee-for-service population and others burdened by significant healthcare expenses.

3 Data

3.1 Data sources

Our dataset merged four sources: the United States Centers for Medicare and Medicaid Services (CMS), the Massachusetts Institute of Technology (MIT) Election Data and Science Lab, the University of Wisconsin (UW) Population Health Institute County Health Rankings & Roadmaps, and the United States Department of Agriculture in conjunction with the United States Census Bureau (Census).

The data containing per capita health expenses by county was pulled from the United States Centers for Medicare and Medicaid Services (CMS). The CMS maintains various public use files that enable researchers and policymakers to evaluate trends in healthcare utilization and spending by geography for the medicare-fee-for-service population.[4] The specific file we extracted data from contains county-level entries from 2007-2019. Only per capita health cost data from 2019 was examined. Given many of the other features were irrelevant in this study or were repeated in the county socioeconomic and health or demographic data all other data from this set was ignored.

The MIT Election Data and Science Lab exists to support advances in election science by collecting, analyzing, and sharing core data and findings.[5] The aforementioned data provided by MIT contains numerous variables specific to the county-level including total voter count, votes per candidate, as well as each candidate's political party affiliation. Only the most recent data was utilized in our examination: the 2020 general presidential election. The major political party affiliation of each winning candidate was considered instead of their name.

The UW data is a compilation of numerous data sources provided as part of the County Health Rankings & Roadmaps program. This University of Wisconsin program serves to promote health equity.[6] Their aim is to build awareness of the multiple factors that influence health. Our aim was to utilize the various socioeconomic, health, environmental, and clinical care data to inform our models. To do this, data from 2019 and earlier

[3] Sylvia Burwell, 'Setting Value-Based Payment Goals — HHS Efforts to Improve U.S. Health Care,' *NHS*, 2015. https://canceradvocacy.org/wp-content/uploads/2015/03/Burwell-NEJM-Value-Based-Payment-Goals.pdf

[4] Centers for Medicare and Medicaid Services: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Geographic-Variation/GV_PUF

[5] MIT Election Data and Science Lab: https://electionlab.mit.edu/about

[6] University of Wisconsin Population Health Institute: https://www.countyhealthrankings.org/about-us

was utilized to create a broader understanding of each county. Of course, then an underlying assumption of our report is that this data changes relatively little year to year and has not been greatly impacted by the COVID-19 pandemic. Future analyses should aim to more consistently include up-to-date information. Explanatory variables that also appeared in the Census data were ignored and instead the corresponding variable was examined from the Census data.

The mission of the Census Bureau is to serve as the nation's leading provider of quality data about the people and the economy of the United States. It operates under Title 13 and Title 26 of the U.S. Code.[7] This dataset, initially complied by the US Department of Agriculture, provides relevant census-gathered information for each county. Data contained in this dataset concerning measures taken prior to 2019 were ignored. Only variables with complete measurements for each observations were considered.

3.2 Data cleaning

Prior to merging the four datasets by FIPS, each dataset was individually tidied. First, all county socioeconomic data from UW was converted to lowercase and explanatory variables repeated in the Census data were removed. Next, variables utilized from the CMS data were renamed according to tidy standards. Units were dropped and the newly named `per_capita_health_cost` variable was converted to a double. All data not relevant to this project was deselected. From there, the MIT data was tidied. Values were made lowercase to meet tidy conventions, the data was filtered to only include measures from the most recent presidential election, `party` was converted to a factor variable, and all data besides FIPS and the political party of each county's 2020 general presidential nominee was disregarded. Finally, the Census data was filtered to contain only the most recent data from 2019 and the four datasets were joined.

3.3 Data description

3.4 Obersvations

Our merged dataset contains 824 observations, each corresponding to a United States's Federal Information Processing Standards identity (FIPS) aligned with a county.

3.5 Response variable

The response variable, `per capita health cost` is pulled from the CMS data. This variable is defined as the cost of all healthcare-related expenses (out-of-pocket medical services, insurance premiums, personal healthcare goods and services) per capita in USD during 2019. This measure is specific to the population subgroup with the highest medical expenses, the medicare fee-for-service population: those 65+ or suffering from disability on Medicare Plan A and/or B.

3.6 Features

From the merged dataset we included 80 explanatory variables in our analysis. For a detailed specification of these variables, refer to **Appendix A**.

3.7 Data allocation

Prior to constructing our predictive models, we removed NA values from the observations in our dataset. Given some data analysis methods we employed require that no variables contain NA fields, we removed NAs for consistency purposes.

We then split our dataset into two subsets: a training dataset used for building our predictive models and a test dataset used for evaluating our models. We used an 80-20 split, such that the training dataset consists of 80% of our observations and the dataset consists of 20% of our observations. Although this train-test split was performed separately for each class of methods, we utilized a random seed to ensure that each split led to the same results.

[7] US Census Bureau: https://www.census.gov/about/what.html

3.8 Data exploration

3.8.1 Response

To understand the response variable's distribution, `per capita health cost` was plotted in **Figure 1**. The data appears to be mostly normal with a few outlines. The median per capita health expense is $10,428.

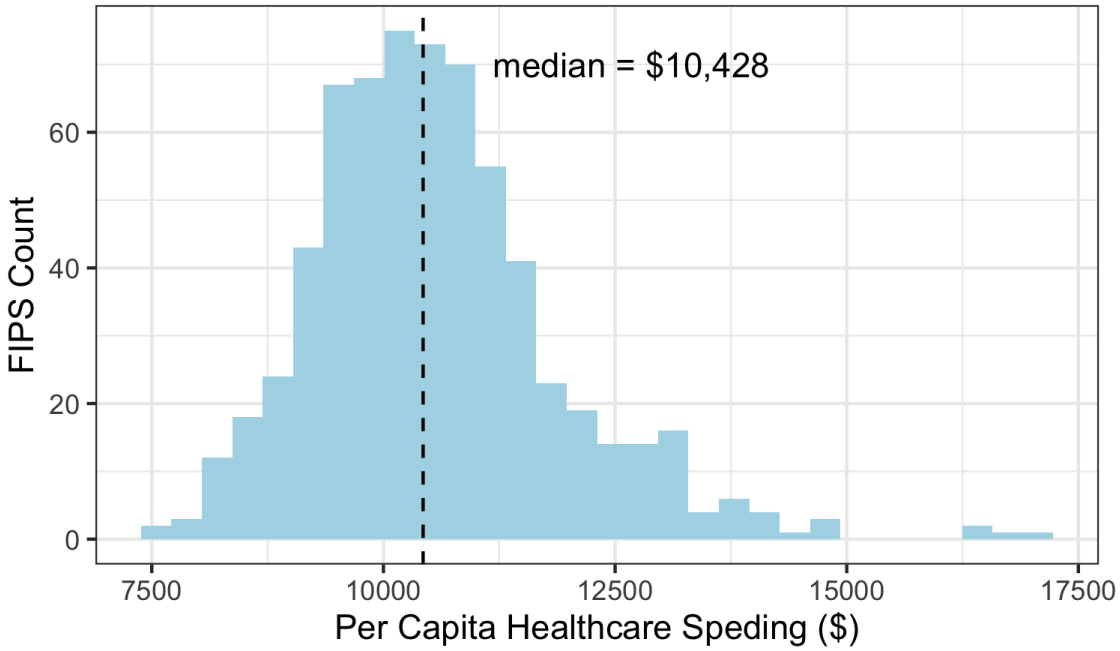

Figure 1: Per Capita Health Cost Distribution

Table 1 shows the FIPS that spend the most on healthcare per capita: Baltimore, MD; Miami-Dade, FL; Los Angeles, CA; Washington, LA; Essex, NJ. **Table 2** displays the FIPS that spend the least per capita on healthcare: Delta, CO; Santa Fe, NM; Tioga, PA; Mesa, CO; Outagamie, WI.

Table 1: FIPS with Highest Per Capita Health Cost

FIPS	Per Capita Spending (USD)
24510	17141
12086	16894
06037	16404
22117	16392
34013	14893

Table 2: FIPS with Lowest Per Capita Health Cost

FIPS	Per Capita Spending (USD)
08029	7624
35049	7653
42117	7714
55087	7945
53029	7980

3.8.2 Features

Figure 2 examines the correlation among our explanatory variables. Correlation among variables varies widely, however, none of the explanatory variables are negatively correlated to any other.

Figure 3 frames the distribution of `per capita health cost` according to political party. Counties who supported the democrat nominee for the 2020 presidential election have a higher cost of healthcare per capita on average. The analysis below will attempt to explain why.

4 Modeling

4.1 Regression-based methods

4.1.1 Ordinary least squares

Given **Figure 4** is roughly normal, a linear regression of `per capita health cost` was run on all 80 explanatory variables. The following variables were significantly associated with the response at the 0.05 level:

- `children_freelunches`
- `housing_overcrowding`
- `homeownership`
- `obesity_pct`
- `partyrepublican`
- `hispanic`
- `household_has_computer`
- `housing_mobile_homes`
- `per_capita_income`
- `poverty`
- `poverty_65_and_older`

See the entire summary statistics in **Appendix.**

4.1.2 Penalized regression

Although the ordinary least squares method worked well, fitting a linear model with eighty explanatory variables may lead to suboptimal predictions given the high variance cost. Aiming to better optimize our models, two cross-validated regressions, where optimal values of lambda were chosen according to the one-standard-error rule, were run: ridge and least absolute shrinkage and selection operator (LASSO).

The ridge trace plot is shown in **Figure 5.** The following features were selected:

- `high_housing_cost`
- `preventable_hospitalization`
- `inactive_perc`
- `pop`

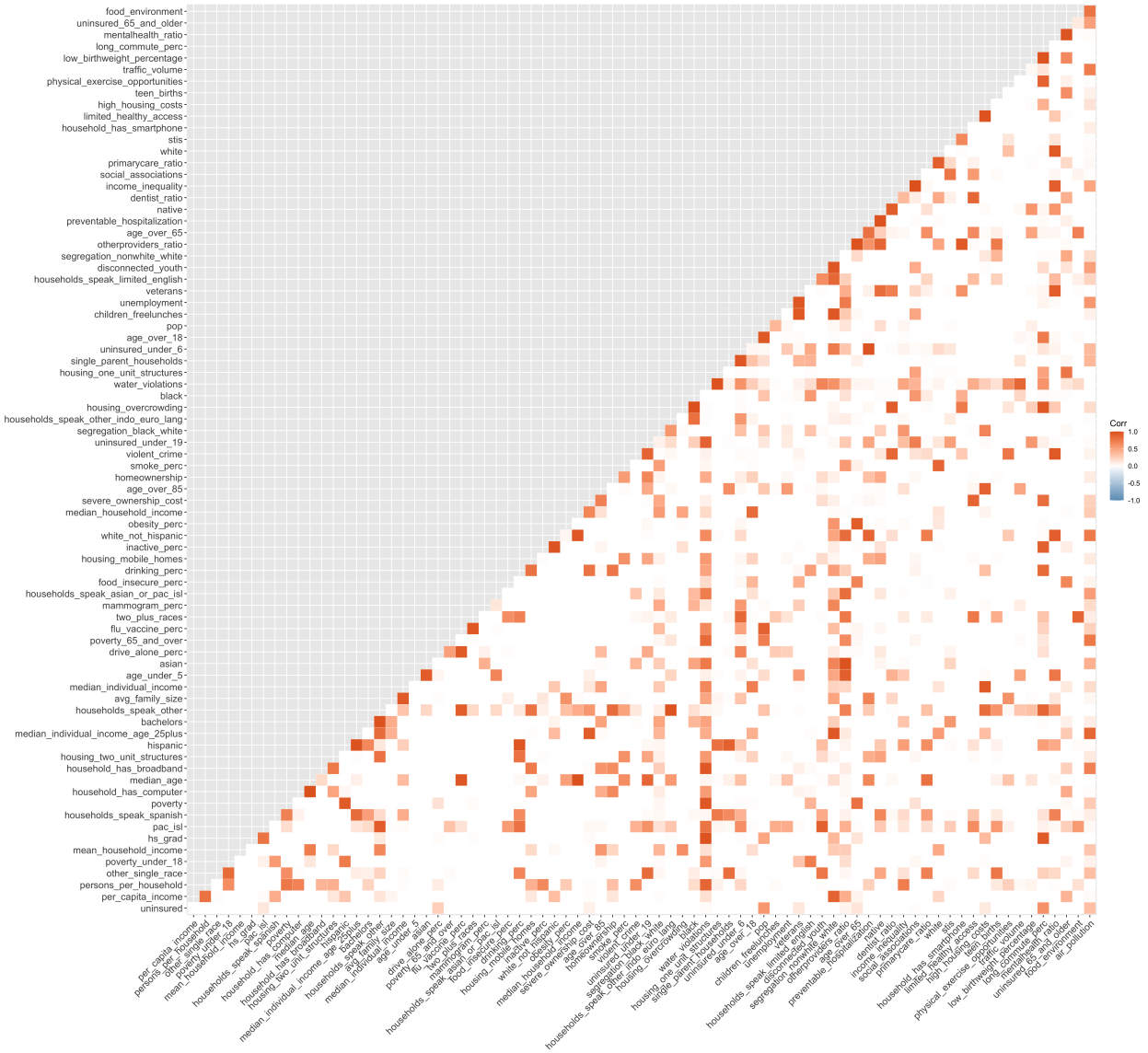

Figure 2: Explanatory Variables Correlation Plot

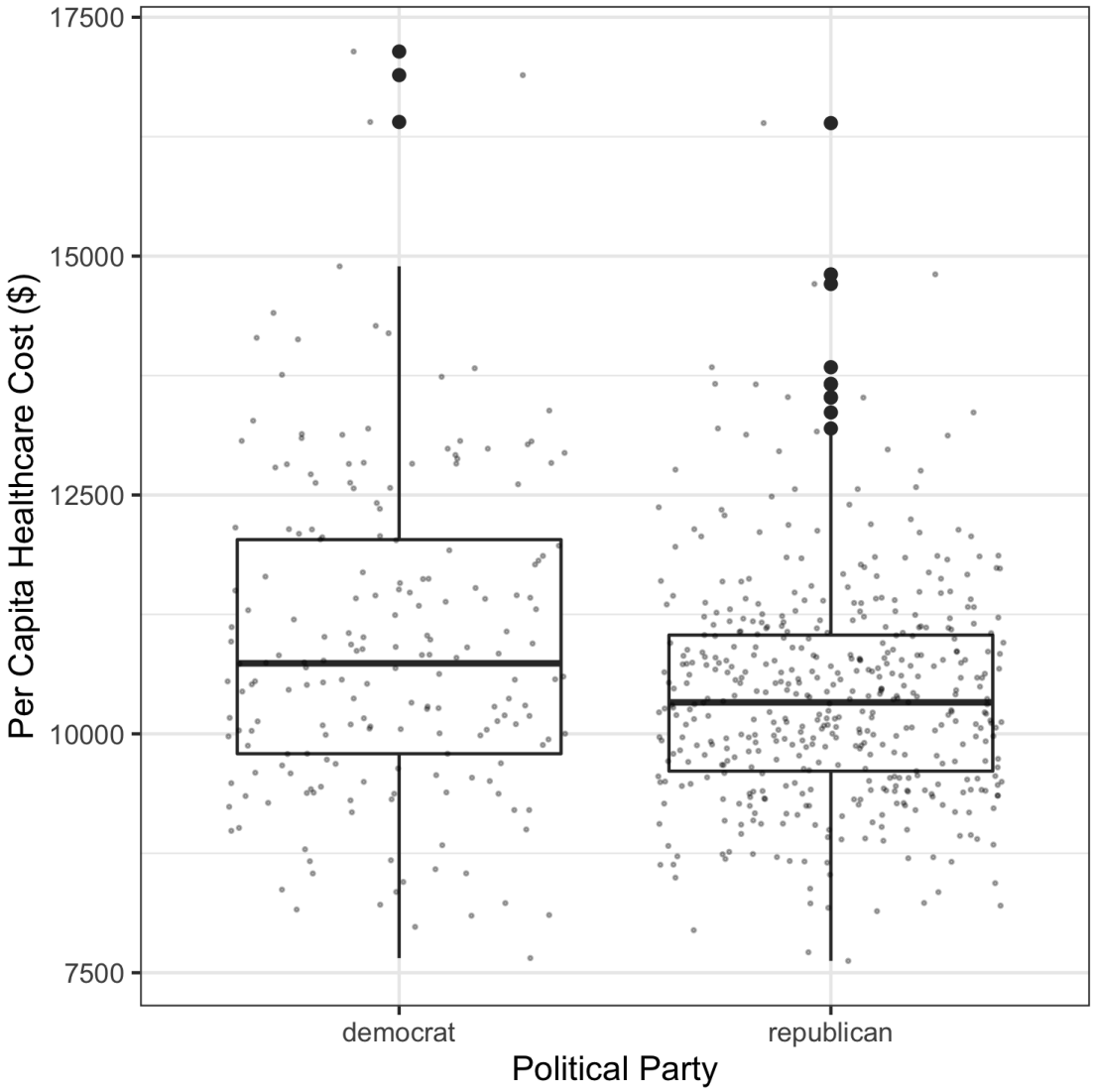

Figure 3: Per Capita Health Cost Distrubution By Political Party.

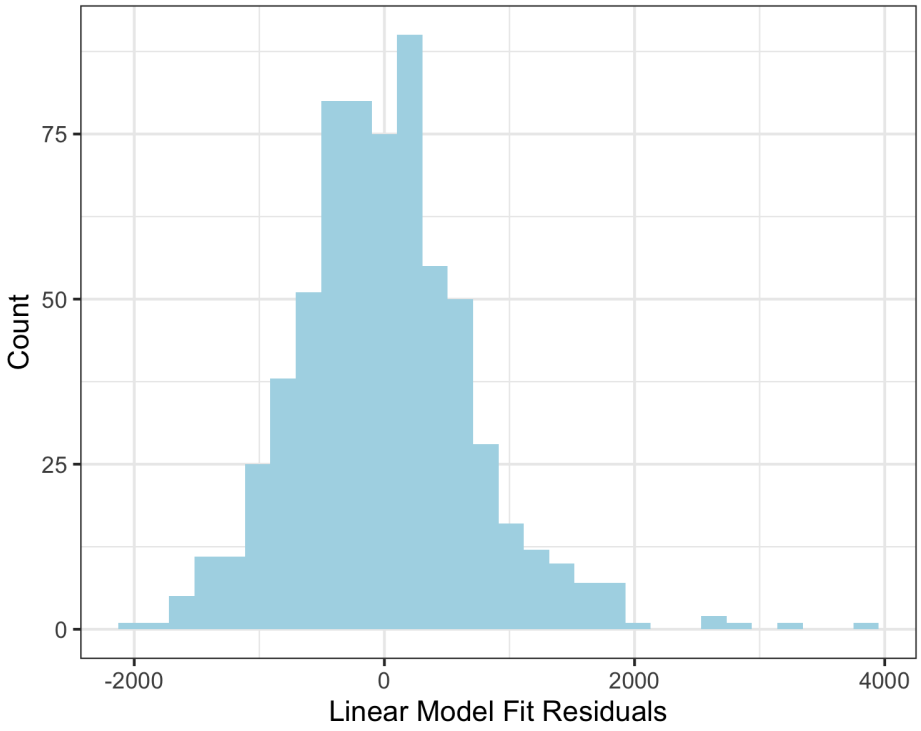

Figure 4: Linear Regression Residuals Plot.

- `long_commute_perc`
- `housing_one_unit_structures`

For lasso, **Figure 6** shows the CV plot, **Figure 7** shows the trace plot, and **Table 3** shows the selected features and their coefficients.

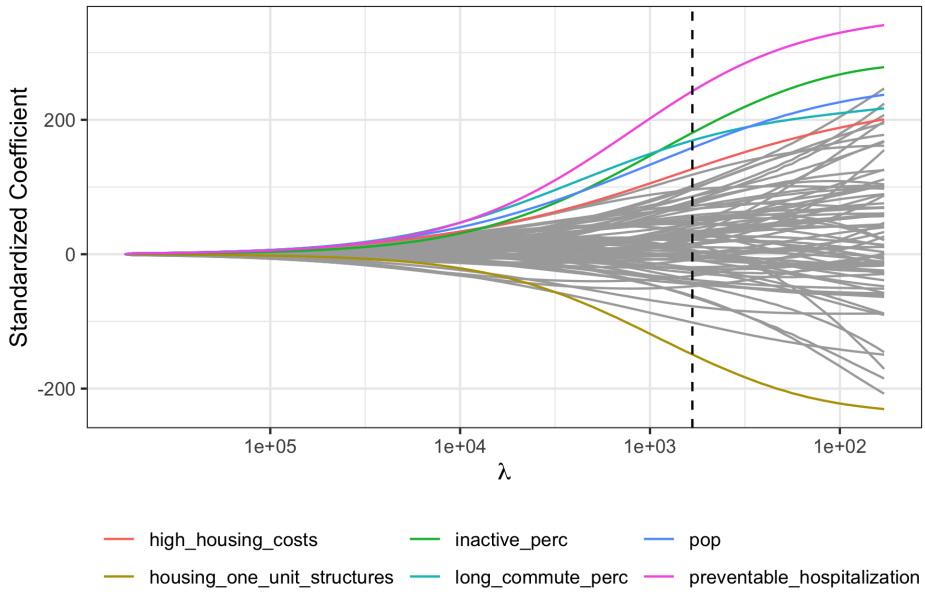

Figure 5: Ridge trace plot.

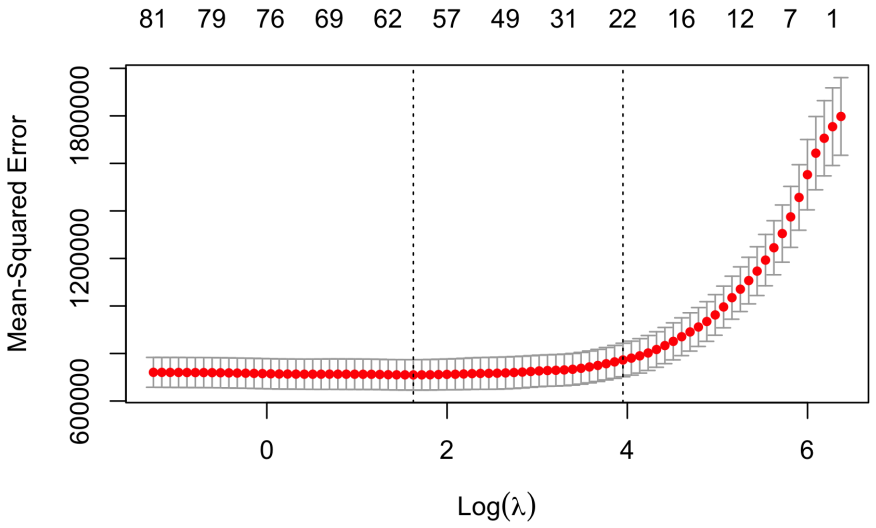

Figure 6: Lasso CV plot.

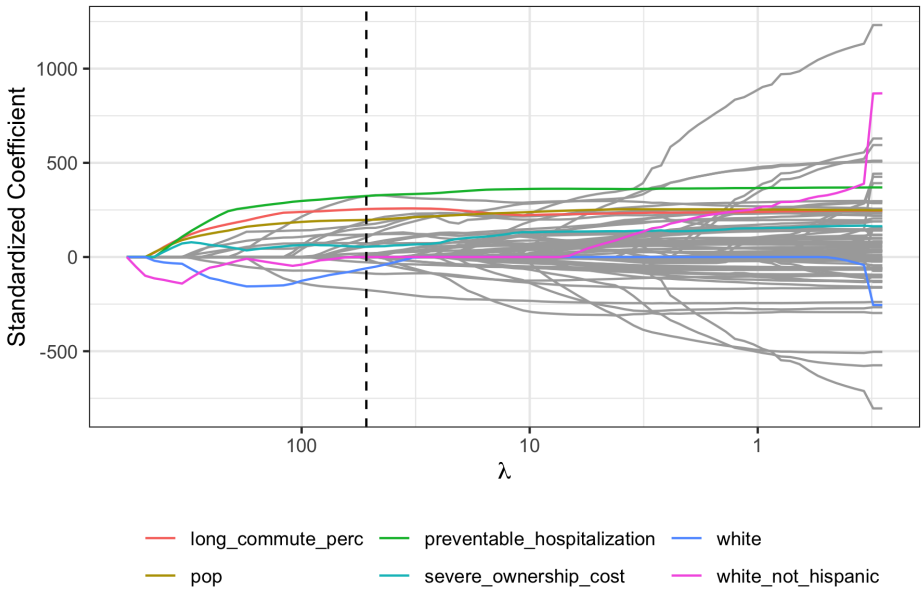

Figure 7: Lasso trace plot.

Table 3: Standardized coefficients for features in the lasso model based on the one-standard-error rule.

Feature	Coefficient
preventable_hospitalization	323.87
inactive_perc	323.25
long_commute_perc	254.70
pop	196.78
single_parent_households	188.66
housing_one_unit_structures	-175.17
high_housing_costs	171.62
mean_household_income	154.17
households_speak_limited_english	119.41
avg_family_size	115.23
median_individual_income_age_25plus	103.49
veterans	-85.79
households_speak_other_indo_euro_lang	73.08
age_over_85	69.64
white	-60.02
two_plus_races	55.98
severe_ownership_cost	54.40
mammogram_perc	-27.91
violent_crime	19.15
low_birthweight_percentage	13.06
partydemocrat	-10.47
uninsured_under_19	2.98

If lambda is chosen according to the one-standard-error rule, six features are selected:

- long_commute_perc
- preventable_hospitalization
- white
- pop
- severe_ownership_cost
- white_not_hispanic

4.2 Tree-based methods

4.2.1 Random forest

Tuning the random forest model on all different possible values of m, the optimal number of features to consider at each tree split, we found the optimal value of m to be 36, as seen in **Figure 8.**

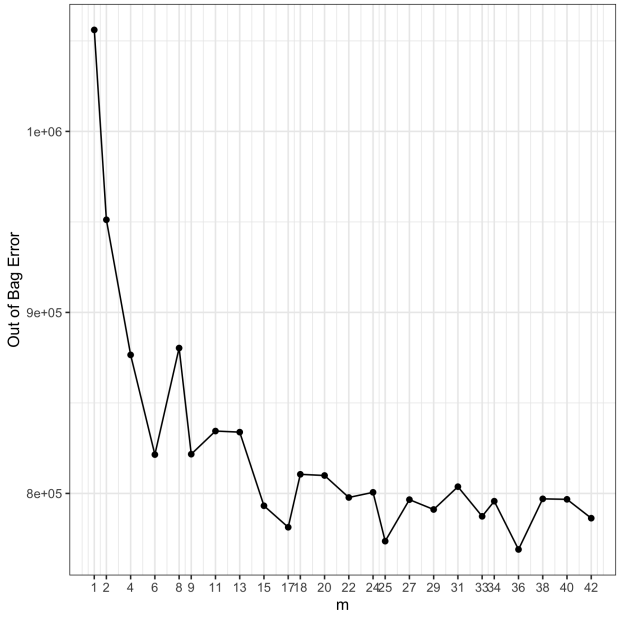

Figure 8: Out of Bag Error Plot.

Figure 9 shows the cross-validated training error sharply decreases as the number of trees increases prior to plateauing around $B = 250$, where B corresponds to the number of fitted trees.

Examining variable importance in our random forest model, we generated **Figure 10.** Out of Bag variable importance estimates by permutation how influential the predictor variables in the model are at predicting the response. Purity based importance is a measure of the degree of improvement in node purity that results from splitting on a given feature.

4.2.2 Boosting

Starting with default parameters of 1000 trees, shrinkage factor = 0.1, interaction depth = 1, and a subsampling fraction of 0.5, we then tuned our boosted model. From **Figure 11 we see that the optimal interaction depth is 2 given this depth attains the miniumum cross-validated error. The optimal number of trees is 246.

From **Table 4** we see the top ten features.

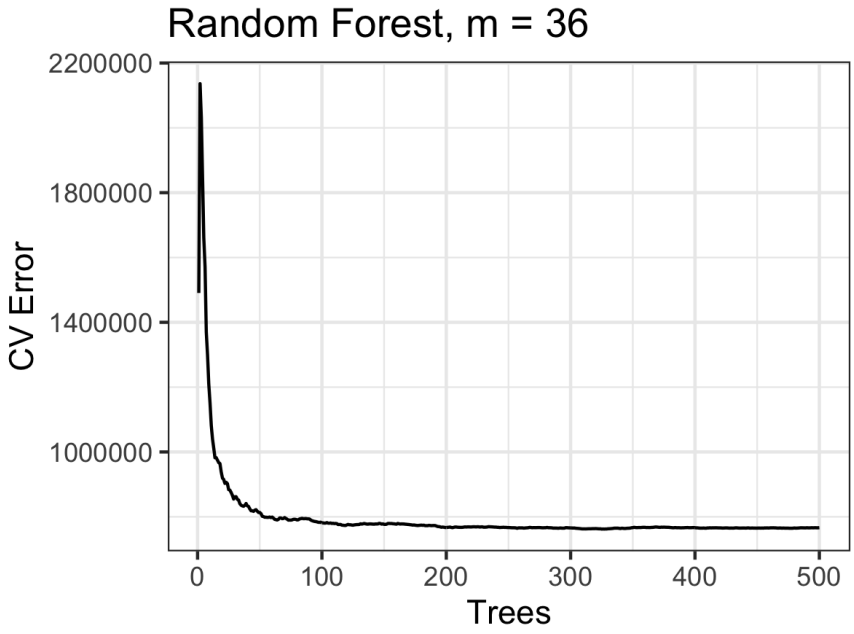

Figure 9: Optimal m CV Error Plot.

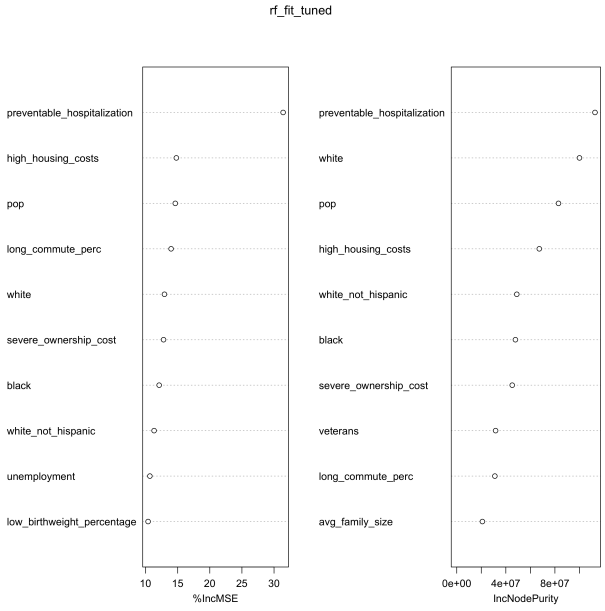

Figure 10: Random Forest Importance Plot.

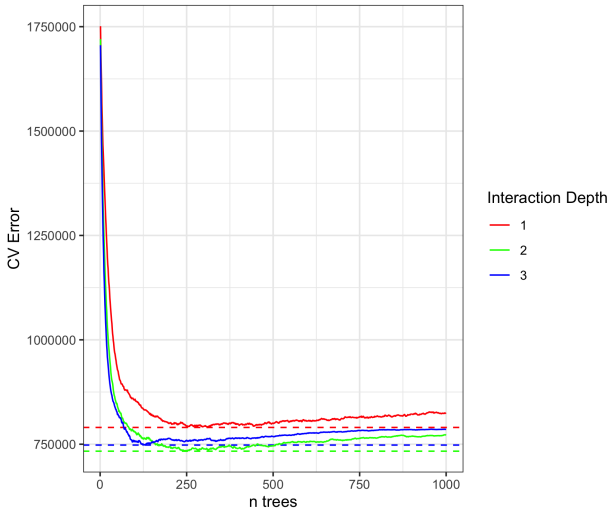

Figure 11: Plot of CV errors against the number of trees for each interaction depth.

Table 4: Optimal Boosting Model Relative Influence.

Variable	Relative Influencce
preventable_hospitalization	11.70
high_housing_costs	8.81
white	6.42
long_commute_perc	5.22
pop	4.77
black	4.64
poverty_65_and_over	3.16
housing_one_unit_structures	3.13
households_speak_other_indo_euro_lang	2.88
households_speak_limited_english	2.52

5 Conclusions

5.1 Method Comparison

Test root-mean-standard prediction errors (RMSE) for all methods considered are shown in **Table 5**.

Table 5: Root-mean-squared prediction errors.

Model	Test RMSE
OLS	720
Boosting	728
Random Forest	751
Ridge	761
Lasso	773

Ordinary least squares had the lowest test RMSE, followed by boosting and random forest. All three of these

methods have high predictive accuracy. Ridge and LASSO, however, were not far off either. Regardless of differences in test RMSE, the majority of methods overlap in their identification of:

- `long_commute_perc`
- `preventable_hospitalization`
- `pop`
- `severe_ownership_cost`
- `white_not_hispanic`

as variables of importance.

5.2 Takeaways

Policy makers, insurance companies, and individuals stand to benefit from our analysis. Our results point to key determinants of health costs. Namely, `long_commute_perc`, `severe_ownership_cost`, and `preventable_hospitalization`. These variables identified across the various regression and tree-based methods are considered robust.

The inclusion of `white` and `white_not_hispanic` as a strong predictor in some of the models points to larger systemic issues here in the United States. Race serves as a proxy for many other variables, and consequentially, reveals not only the problems with racial discrimination in this country but also the statistical problem of confounds.

Understanding which socioeconomic variables affect individuals' healthcare spending within the medicare-fee-for-service population will help inform individuals' personal financial decisions when choosing an appropriate healthcare plan. Accurate predictive models are also critical to efficiently target government relief and aid programs centered on this population. Knowing that someone who lives in a county with a large percentage of the population enduring long daily commutes or one with a high percentage of households that spend 50% or more of their household income on housing can guide policymakers' when reforming medicare legislation to ensure that those with the highest need (i.e. those with the highest healthcare-related expenses) benefit regardless of their current access to government welfare services. Knowing that someone lives in a county with a high percentage of preventable hospitalizations, however, may also result in insurance companies hiking premium's in that area given the county's higher cost of care per capita. Whether or not it is ethical to charge someone more for an essential service simply because of the neighborhood they live in remains a question to be answered and is beyond the scope of this report.

5.3 Limitations

5.3.1 Dataset limitations

Given both the CMS and UW dataset contain measures more than two years old, it is possible that on account of COVID-19 and other changes, this data does not accurately reflect the current make-up of county socioeconomic factors corresponding to our per capita health cost figures, which may have also been affected by the pandemic. Correlation among our explanatory variables suggests that some of these variables may have been confounds which only distorted our analysis. Only FIPS with data corresponding to all 82 variables were analyzed. Only 824 of the 3,241 possible observations by FIPS were made given the need to merge the datasets and remove any observations containing NA values. Future iterations of this study should attempt to avoid this limitation.

5.3.2 Analysis limitations

Variables selected in the LASSO regression as well as variables found to be important in the tree-based methods might be misleading in that, given how variable selection works, it's possible that some selected variables are simply representative of a larger group of correlated variables. Additionally, while splitting the data into training and testing datasets allows for a more unbiased test of the models, it is possible that had a different random seed been used when splitting the data, the p-values derived in the OLS regression, the selected variables in the penalized regression methods, and the variables marked important in the tree-based

methods might have been different. Our results may have been quite different had we incorporated a different set of variables too. Differences in the genetic make-up of a county's population, for example, are highly difficult to take into account. People move, family lineages are not always accurately recorded, and disease-science has really only taken hold in the last century. Our understanding of many of the key ways genetics affect health outcomes and consequently an individual's healthcare expenses is still unknown.

5.4 Future Directions

To build on our report, future iterations of this study ought to include data only from the current year or perhaps look at how these variables change over time within a county and affect per capita healthcare costs over an individuals lifetime. Future studies should aim to be more inclusive in their analysis of all 3,241 FIPS in the United States in order to build a more accurate predictive model. Examining data from other countries may also enhance our models and understadning.

A Appendix

A.1 Descriptions of features

- `pop`: population.
- `white`: percent of population that is white alone.
- `black`: percent of population that is black alone.
- `native`: percent of population that is Native American alone.
- `asian`: percent of population that is Asian alone.
- `pac_isl`: percent of population that is Native Hawaiian or other Pacific Islander alone.
- `other_single_race`: percent of population that is some other race alone.
- `two_plus_races`: percent of population that is two or more races.
- `hispanic`: percent of population that identifies as Hispanic or Latino.
- `white_not_hispanic`: percent of population that is white alone, not Hispanic or Latino.
- `median_age`: median age.
- `age_under_5`: percent of population under 5.
- `age_over_85`: percent of population 85 and over.
- `age_over_18`: percent of population 18 and over.
- `age_over_66`: percent of population 65 and over.
- `mean_work_travel`: mean travel time to work.
- `persons_per_household`: persons per household.
- `avg_family_size`: average family size.
- `housing_one_unit_structures`: percent of housing units in 1-unit structures .
- `housing_two_unit_structures`: percent of housing units in multi-unit structures .
- `housing_mobile_homes`: percent of housing units in mobile homes and other types of units .
- `median_individual_income_age_25plus`: median individual income.
- `hs_grad`: percent of population 25 and older that is a high school graduate.
- `bachelors`: percent of population 25 and older that earned a Bachelor's degree or higher.
- `households`: total households.
- `households_speak_spanish`: percent of households speaking Spanish.
- `households_speak_other_indo_euro_lang`: percent of households speaking other Indo-European language.
- `households_speak_asian_or_pac_isl:` percent of households speaking Asian and Pacific Island language.
- `households_speak_other`: percent of households speaking non European or Asian/Pacific Island language.
- `households_speak_limited_english`: percent of limited English-speaking households.
- `poverty`: percent of population below the poverty level.
- `poverty_under_18`: percent of population under 18 below the poverty level.
- `poverty_65_and_over`: percent of population 65 and over below the poverty level.

- `mean_household_income`: mean household income.
- `per_capita_income`: per capita money income in past 12 months.
- `median_household_income`: median household income.
- `veterans`: percent among civilian population 18 and over that are veterans.
- `unemployment_rate`: unemployment rate among those ages 20-64.
- `uninsured`:percent of civilian non-institutionalized population that is uninsured.
- `uninsured_under_6`: percent of population under 6 years that is uninsured.
- `uninsured_under_19`: percent of population under 19 that is uninsured.
- `uninsured_65_and_older`: percent of population 65 and older that is uninsured.
- `household_has_computer`: percent of households that have desktop or laptop computer.
- `household_has_smartphone`: percent of households that have smartphone.
- `household_has_broadband`: percent of households that have broadband internet subscription.
- `party`: political party affiliation of the winning presidential candidate in the 2020 general election.
- `smoke_perc`: percentage of adults who are current smokers.
- `obesity_perc`: percentage of the adult population (age 20 and older) reporting a body mass index (BMI) greater than or equal to 30 kg/m2.
- `food_environment`: index of factors that contribute to a healthy food environment, from 0 (worst) to 10 (best).
- `inactive_perc`: percentage of adults age 20 and over reporting no leisure-time physical activity.
- `physical_exercise_opportunities`: percentage of population with adequate access to locations for physical activity
- `food_insecure_perc`: percentage of population who lack adequate access to food.
- `limited_healthy_access`: percentage of population who are low-income and do not live close to a grocery store.
- `drinking_perc`: percentage of adults reporting binge or heavy drinking.
- `stis`: number of newly diagnosed chlamydia cases per 100,000 population.
- `teen_births`: number of births per 1,000 female population ages 15-19.
- `low_birthweight_percentage`: percentage of live births with low birthweight ($<$ 2,500 grams).
- `primarycare_ratio`: ratio of population to primary care physicians.
- `dentist_ratio`: ratio of population to dentists.
- `mentalhealth_ratio`: ratio of population to mental health providers.
- `otherproviders_ratio`: ratio of population to primary care providers other than physicians.
- `preventable_hospitalization`: rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees.
- `mammogram_perc`: percentage of female Medicare enrollees ages 65-74 that received an annual mammography screening.
- `flu_vaccine_perc`: percentage of fee-for-service (FFS) Medicare enrollees that had an annual flu vaccination.
- `disconnected_youth`: percentage of teens and young adults ages 16-19 who are neither working nor in school.
- `unemployment`: percentage of population ages 16 and older who are unemployed but seeking work.
- `income_inequality`: ratio of household income at the 80th percentile to income at the 20th percentile.
- `children_freelunches`: percentage of children enrolled in public schools that are eligible for free or reduced price lunch.
- `single_parent_households`: percentage of children that live in a household headed by a single parent.
- `social_associations`: number of membership associations per 10,000 residents.
- `segregation_black_white`: index of dissimilarity where higher values indicate greater residential segregation between Black and White county residents.
- `segregation_nonwhite_white`: index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents.
- `violent_crime` number of reported violent crime offenses per 100,000 residents.
- `air_pollution`: average daily density of fine particulate matter in micrograms per cubic meter (PM2.5).
- `water_violations`: indicator of the presence of health-related drinking water violations. 1 indicates the presence of a violation, 0 indicates no violation.

- `housing_overcrowding`: percentage of households with overcrowding,
- `high_housing_costs`: percentage of households with high housing costs
- `driving_alone_perc`: percentage of the workforce that drives alone to work.
- `long_commute_perc`: among workers who commute in their car alone, the percentage that commute more than 30 minutes.
- `traffic_volume`: average traffic volume per meter of major roadways in the county.
- `homeownership`: percentage of occupied housing units that are owned.
- `severe_ownership_cost`: percentage of households that spend 50% or more of their household income on housing.

A.2 Linear regression summary statistics

```
Call:
lm(formula = per_capita_health_cost ~ . - fips, data = health_train)

Residuals:
   Min     1Q Median     3Q    Max
 -1949   -446    -44    376   3928

Coefficients: (1 not defined because of singularities)
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               1.36e+04   4.38e+04    0.31  0.75526
low_birthweight_percentage                7.16e+03   4.65e+03    1.54  0.12376
food_environment                          6.27e+02   5.35e+02    1.17  0.24131
physical_exercise_opportunities           4.71e+02   3.52e+02    1.34  0.18212
teen_births                              -1.08e+04   8.22e+03   -1.31  0.18944
limited_healthy_access                    4.97e+03   4.80e+03    1.04  0.30106
stis                                     -1.15e+04   3.59e+04   -0.32  0.74888
primarycare_ratio                        -2.47e-03   4.86e-02   -0.05  0.95951
dentist_ratio                            -2.18e-02   4.27e-02   -0.51  0.61027
mentalhealth_ratio                        5.36e-02   4.25e-02    1.26  0.20795
otherproviders_ratio                     -9.51e-02   6.84e-02   -1.39  0.16508
disconnected_youth                        1.85e+03   1.27e+03    1.46  0.14567
unemployment                              6.87e+03   3.95e+03    1.74  0.08274 .
income_inequality                        -4.13e+00   1.29e+02   -0.03  0.97442
children_freelunches                      8.67e+02   4.08e+02    2.13  0.03397 *
single_parent_households                  2.89e+03   1.07e+03    2.69  0.00729 **
social_associations                       2.29e+05   1.47e+05    1.56  0.11945
water_violations                         -1.36e+02   7.02e+01   -1.94  0.05327 .
high_housing_costs                        6.04e+03   4.19e+03    1.44  0.15034
housing_overcrowding                      9.53e+03   4.35e+03    2.19  0.02862 *
segregation_black_white                  -4.01e+00   3.88e+00   -1.03  0.30220
segregation_nonwhite_white                2.32e+00   4.73e+00    0.49  0.62391
homeownership                             9.85e+03   4.65e+03    2.12  0.03456 *
severe_ownership_cost                     4.62e+03   4.46e+03    1.04  0.30069
smoke_perc                                2.87e+03   2.34e+03    1.22  0.22194
obesity_perc                              2.72e+03   1.20e+03    2.27  0.02355 *
inactive_perc                             4.17e+03   1.23e+03    3.39  0.00076 ***
drinking_perc                             9.82e+02   1.78e+03    0.55  0.58193
food_insecure_perc                        1.51e+04   9.97e+03    1.52  0.13014
preventable_hospitalization               2.29e-01   2.84e-02    8.07  4.2e-15 ***
mammogram_perc                           -1.58e+03   8.70e+02   -1.82  0.06969 .
flu_vaccine_perc                          1.67e+02   7.81e+02    0.21  0.83070
violent_crime                             3.93e-01   2.44e-01    1.61  0.10801
air_pollution                            -3.31e+01   2.56e+01   -1.29  0.19654
drive_alone_perc                          1.32e+03   9.75e+02    1.35  0.17685
long_commute_perc                         2.25e+03   5.57e+02    4.04  6.0e-05 ***
traffic_volume                           -2.05e-01   1.43e-01   -1.43  0.15322
partyrepublican                           2.19e+02   1.07e+02    2.05  0.04108 *
age_over_18                               7.66e+01   4.86e+01    1.58  0.11519
age_over_65                               1.39e+02   3.57e+01    3.89  0.00011 ***
age_over_85                               4.30e+01   1.20e+02    0.36  0.71980
age_under_5                               4.40e+01   1.47e+02    0.30  0.76413
asian                                    -4.14e+02   4.39e+02   -0.94  0.34552
avg_family_size                           8.86e+02   5.71e+02    1.55  0.12079
bachelors                                -2.41e+01   1.29e+01   -1.86  0.06293 .
black                                    -4.18e+02   4.36e+02   -0.96  0.33747
hispanic                                  1.86e+02   7.54e+01    2.47  0.01383 *
household_has_broadband                  -1.72e+01   1.36e+01   -1.27  0.20547
household_has_computer                   -3.36e+01   1.66e+01   -2.03  0.04296 *
household_has_smartphone                  3.55e+01   1.32e+01    2.70  0.00723 **
households_speak_asian_or_pac_isl        -4.52e+01   5.93e+01   -0.76  0.44565
households_speak_limited_english          1.18e+02   4.42e+01    2.67  0.00784 **
households_speak_other                   -2.28e+01   2.30e+01   -0.99  0.32203
households_speak_other_indo_euro_lang     2.50e+01   1.55e+01    1.61  0.10793
households_speak_spanish                 -1.16e+01   2.89e+01   -0.40  0.68894
housing_mobile_homes                      9.37e+01   4.59e+01    2.04  0.04165 *
housing_one_unit_structures              -2.93e+01   8.42e+00   -3.48  0.00053 ***
housing_two_unit_structures                     NA         NA      NA       NA
hs_grad                                   5.50e+01   2.05e+01    2.68  0.00767 **
mean_household_income                     6.81e-02   2.39e-02    2.85  0.00448 **
median_age                               -1.19e+02   3.80e+01   -3.13  0.00186 **
median_household_income                   6.42e-03   2.13e-02    0.30  0.76305
median_individual_income                 -1.04e-02   4.28e-02   -0.24  0.80711
median_individual_income_age_25plus       7.10e-02   2.63e-02    2.70  0.00716 **
native                                   -4.16e+02   4.35e+02   -0.96  0.33951
other_single_race                        -6.11e+02   4.40e+02   -1.39  0.16482
pac_isl                                  -7.36e+02   4.53e+02   -1.63  0.10462
per_capita_income                        -1.26e-01   5.65e-02   -2.22  0.02666 *
persons_per_household                    -9.10e+02   9.18e+02   -0.99  0.32222
pop                                       4.58e-04   8.22e-05    5.57  3.9e-08 ***
poverty                                  -9.57e+01   3.98e+01   -2.40  0.01649 *
poverty_65_and_over                       5.42e+01   2.31e+01    2.34  0.01953 *
poverty_under_18                          3.05e+01   1.87e+01    1.63  0.10267
two_plus_races                           -3.33e+02   4.37e+02   -0.76  0.44601
uninsured                                 3.85e+01   2.48e+01    1.55  0.12159
uninsured_65_and_older                   -1.70e+01   9.90e+01   -0.17  0.86351
uninsured_under_19                        1.59e+01   3.75e+01    0.42  0.67122
uninsured_under_6                        -1.04e+01   2.58e+01   -0.40  0.68801
veterans                                 -6.13e+01   2.14e+01   -2.86  0.00441 **
white                                    -6.05e+02   4.40e+02   -1.38  0.16962
white_not_hispanic                        2.07e+02   7.98e+01    2.60  0.00963 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 768 on 579 degrees of freedom
Multiple R-squared:  0.712,     Adjusted R-squared:  0.672
F-statistic: 18.1 on 79 and 579 DF,  p-value: <2e-16
```

Figure 12: Linear Regression Summary Statistics