Appendix

Data exploration and cleaning

We first examined variables for any skewness, and then tested associations between covariates to identify potential sources of collinearity. We found that several sets of variables were correlated. For those groups of correlated covariates, we generally chose to include the single covariate that was most highly associated with the outcome, cancer mortality rate.

We saw that median age was highly skewed, to the point that we think there is some sort of data error. Keep this in mind for later; for now we remove from the list of selected parameters. We opt instead to use the median age in males as a proxy, since the distribution (discarding the outliers in median age) is comparable to median age overall and median age in females. It's also interesting to note that median age between males and females are highly correlated eith each other, but not much correlated with target death rate. In fact, median age in males is negatively correlated with death rate, while median age for females is positively correlated.

We looked again at the variables for the percent of the population of a certain race and found that log transformation was bimodal. It might make sense to make this binary (i.e. >5% of a certain race or not). In the end, we chose to make a single variable measuring the percent of the population that is non-white (considered minority). This variable was skewed, but we did not perform a transformation for the sake of interpretability.

We combined several variables to create aggregate measures. We initially thought to group counties by region (Northeast, South, Midwest, West), but found that state-level groupings provided better fit and predictive capability. We designated the District of Columbia as being part of Maryland, since it had a single observation. We also created single measures for the percent of the population with a bachelor's degree and percent of the population with health insurance.

Median income and the percent of the population in poverty are correlated with each other as well as the percent of the population with a college degree. Percent in poverty had a slightly higher association with the death rate, and a slightly lower correlation with education (as measured), so we opted to include it over median income.

How do we interpret the lack of association between the number of new cancer cases and incidence rate?

Measure	Min	1st Q	Mean	Median	3rd Q	Max	Std Dev
Avg household size	0.022	2.37	2.48	2.5	2.63	3.97	0.43
Cancer incidence	201.3	420.3	448.27	453.55	480.85	1206.9	54.56
Percent change in number of new cases	1.79	4.33	5.32	5.14	6.25	10.55	1.43
Median age among men	22.4	36.35	39.57	39.6	42.5	64.7	5.23
Pop. percentage with college degree	2.7	13.2	19.44	17.7	23.75	86.3	8.88
Pop. percentage with insurance	65.4	96.25	100.61	101.3	105.8	131.7	7.39
Pop. percentage non-white	0	4.55	16.35	9.94	22.7	89.8	16.38
Unemployment rate	0.4	5.5	7.85	7.6	9.7	29.4	3.45
Pop. percentage married	23.1	47.75	51.77	52.4	56.4	72.5	6.9
Pop. percentage in poverty	3.2	12.15	16.88	15.9	20.4	47.4	6.41
Avg number of clinical studies	0	0	155.4	0	83.65	9762.31	529.63
Cancer mortality rate (per 100,000)	59.7	161.2	178.66	178.1	195.2	362.8	27.75

Model selection

Stepwise

Lasso Model Selection / checking our stepwise coefficients

When we test our original selected variables with a Lasso model, we find the same parameters are down-weighted/removed.

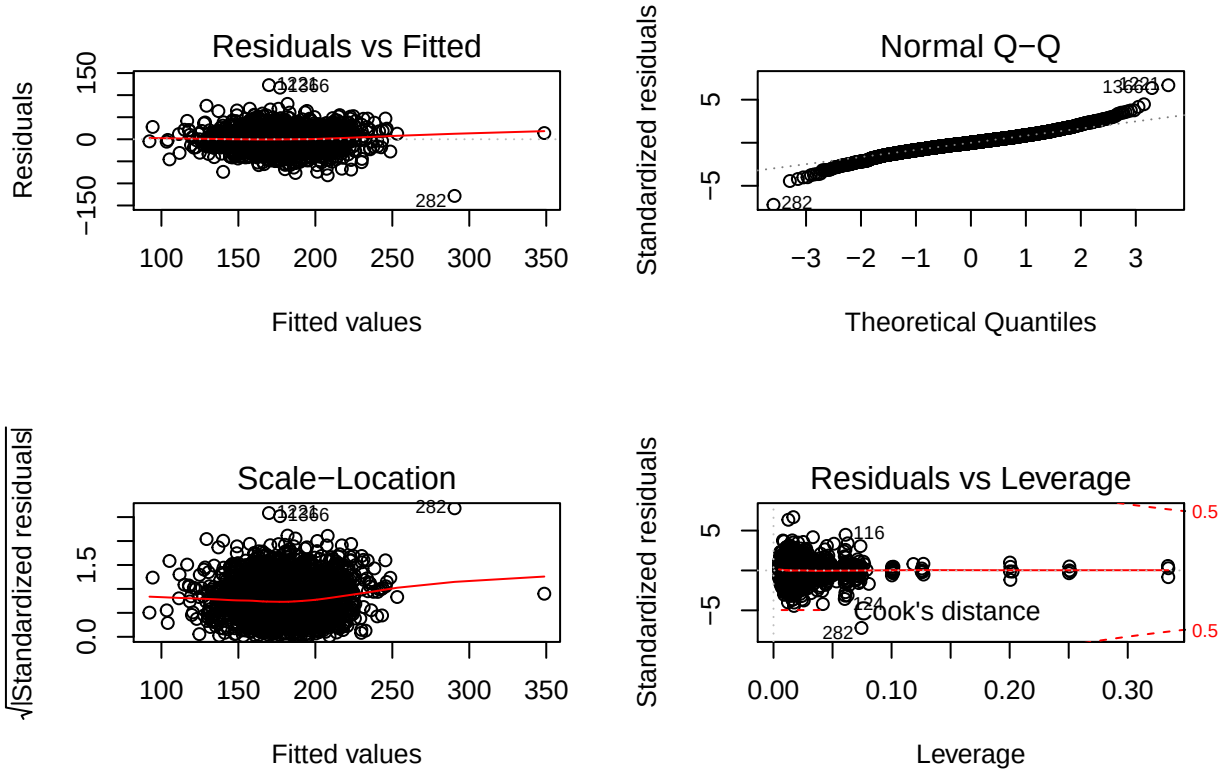
We also tried fitting Lasso on the full original dataset (with our transformed/grouped variables) and found that many of the covariates appeared meaningful. However, this model was seen to be overfitting the data, especially since it chose to include some predictors with known correlation, suggesting presence of multicollinearity.

The stepwise process eliminated the percent of the population with health insurance as a predictor of cancer mortality per capita. However, we suspect this may be a practically meaningful covariate, since it has been shown that those with health insurance tend to have improved cancer outcomes over those without health insurance. This was done on an individual level (https://www.ncbi.nlm.nih.gov/pubmed/29192307) and by relating individual insurance coverage with communities (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5360496/). [Are we at risk of an atomistic fallacy?]

Our final model uses cancer incidence, percent change in the number of new diagnoses from year to year, percent of the population in poverty, median age, percent of the population that is married, percent of the population with a college degree, percent of the population that is non-white, percent unemployment, and the percent of the population with health insurance to predict the average cancer death rate per 100,000 across states.

Model diagnostics

Testing assumptions for final model:



We see that there are long tails on both sides of the distribution. This may be due to outliers at either extreme, so next we examine presence of outliers and potentially influential points.

Skewness still exists without these influential points, but the model's R2 and Adj R2 both improve slightly.

What is the takeaway?

Model validation

K-fold CV

MSE	RMSE	SE
350.0434	18.70943	0.126071

Our 10-fold cross validation shows root MSE of ~18.7. Performing leave-one-out (N-fold) cross validation gives a comparable value. We also examine other criteria (SSE, R-squared, and PRESS):

Criterion comparison

SSE	PRESS	RMSE	Adjusted Rsq
1024772	1068536	18.51924	0.5546789

Commentary