Predicting the change in insurance rate post Affordable Care Act - A supervised learning approach

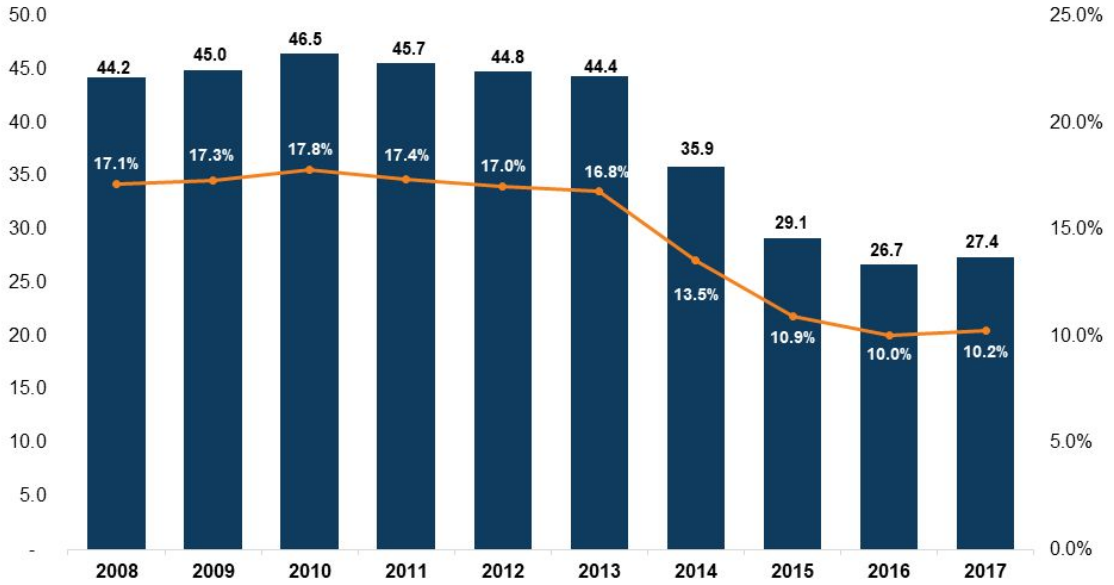
Seán McManus, PhD

KFF
HENRY J KAISER FAMILY FOUNDATION

THINKFUL

Background..

of uninsured individuals and % uninsured:



Why should we care :

- Uninsured cancer patients generally have poorer outcomes and are more likely to die prematurely than persons with insurance, largely because of delayed diagnosis. Results in higher morbidity and higher mortality.
- Uninsured adults with chronic diseases are less likely to receive appropriate care to manage their health conditions than are those who have health insurance. Results in higher morbidity and higher mortality.
- Uninsured adults with hypertension or high blood cholesterol have diminished access to care - higher morbidity and mortality
- Uninsured persons with diabetes are less likely to receive recommended services - higher morbidity and mortality.
- Adults with health insurance that covers any mental health treatment are more likely to receive mental health services and care consistent with clinical practice guidelines.
- Uninsured patients with acute cardiovascular disease are less likely to be admitted to a hospital - higher morbidity and mortality.

A moral issue….

A fiscal responsibility issue….

The hidden risks....

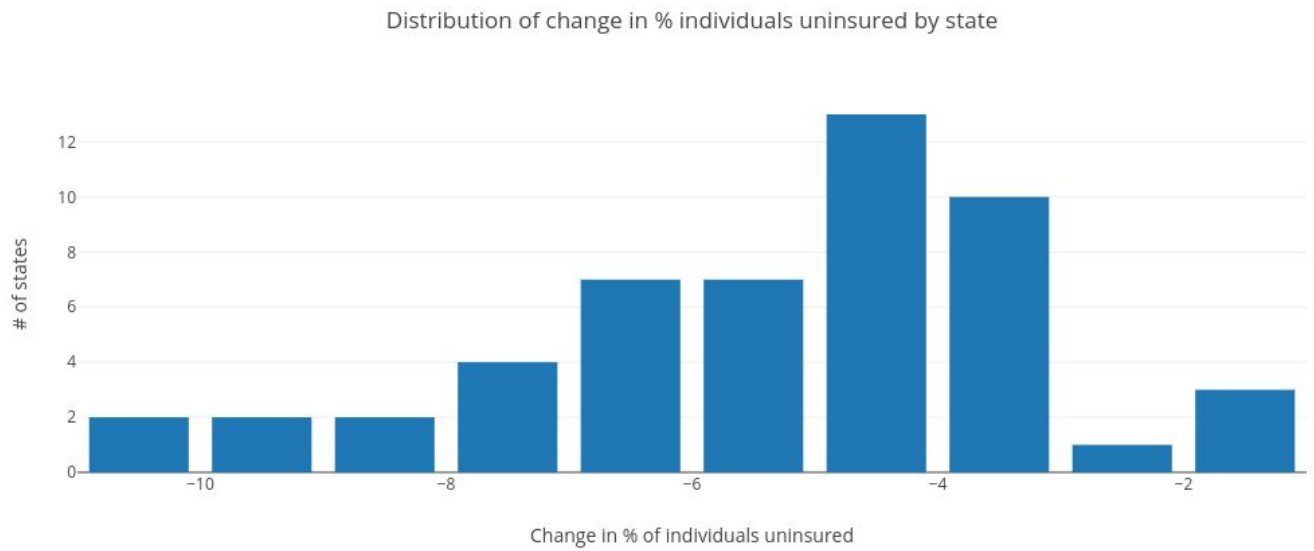
- Contagian of disease
- Loss of productivity and contribution to the economy by the worker
- Loss of productivity and contribution to the economy due to burden on non-professional family care givers (brother taking care of sister, parent taking care of grown adult = potential loss of two productive tax payers)

The money spent on medicaid expansion has had a multiplier effect estimated at 1.5 to 2 times the investment.

In other words, the money invested in decreasing the percentage of individuals uninsured, is effectively a money maker.

[1]

Our problem:



The ACA lead to 20 million people gaining insurance, but the effect of this piece of legislation varied greatly by state.

The question(s):

1. Can we accurately predict the effect size of the ACA , as measured by change in % individuals uninsured, using feature variables that describe that state?
2. Can we establish the metrics that have the biggest effect on the change in % individuals uninsured?

The benefits:

Objective benefits:
Potential to inform strategically crafted legislation that maximizes effect size.
Non-objective benefits:
..this is of major political capital.

The data:

Personally collated data set..from data provided by:

Department of Health and Human Services

and

The Henry J Kaiser Family Foundation

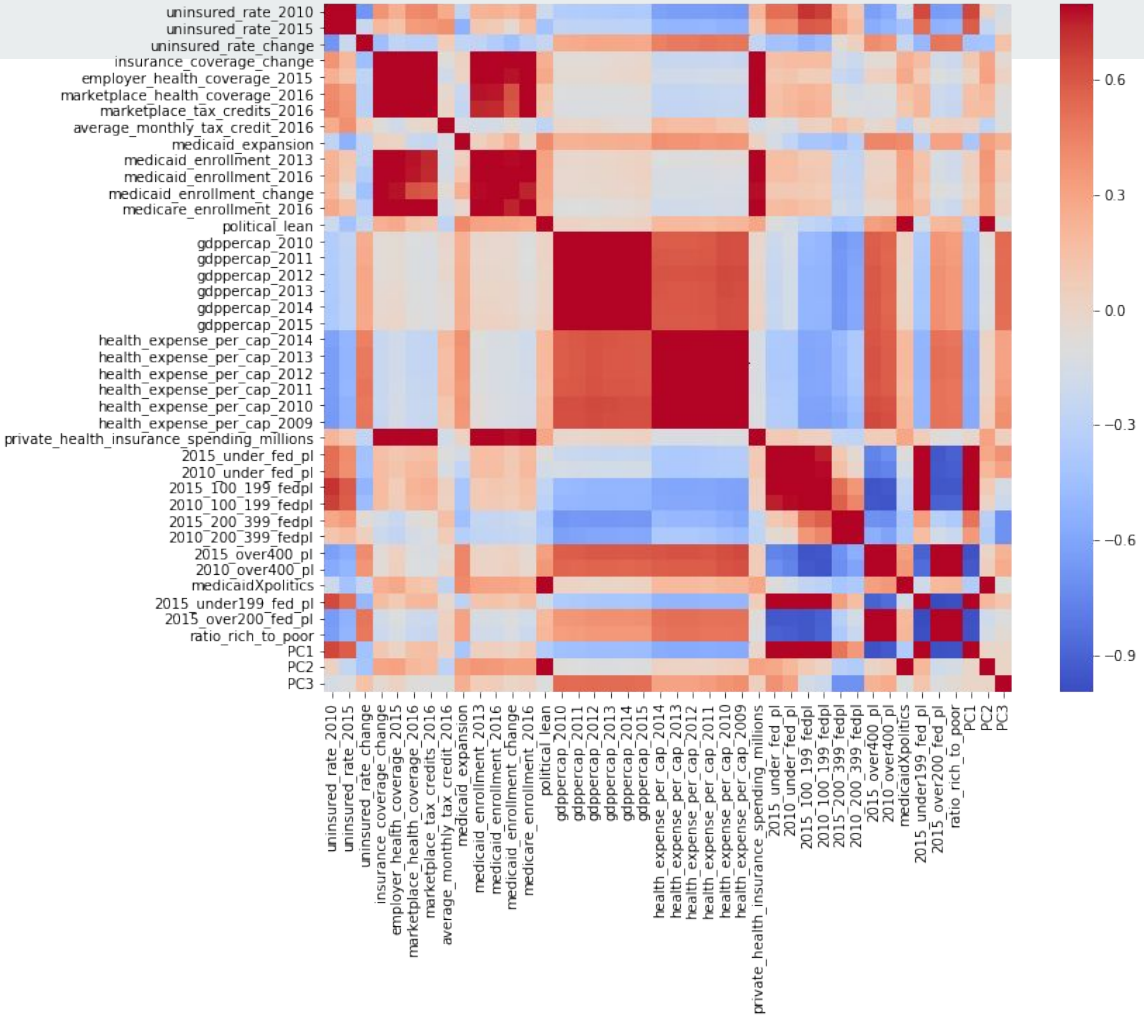
The data - described:

Target feature:

Dataframe shape: 52 samples/rows and 34 features

- Significant feature engineering and subsequent feature selection
- Significant sample size problems
- No missing data

Lot's of descriptive value in the features, perhaps not enough samples..



First task: reduce variables

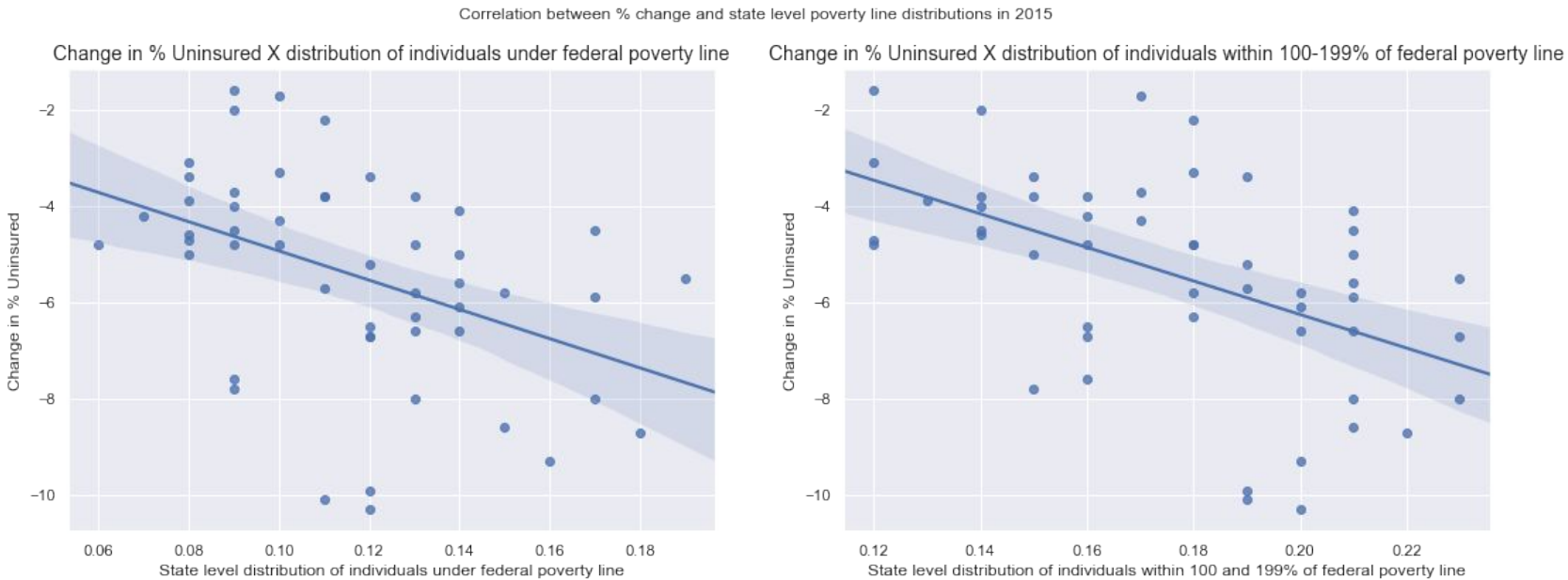
Tried PCA to compress effect of politics and levels of poverty into 1 or 2 variables - ultimately didn't include these.

Process of elimination led to selection of the following:

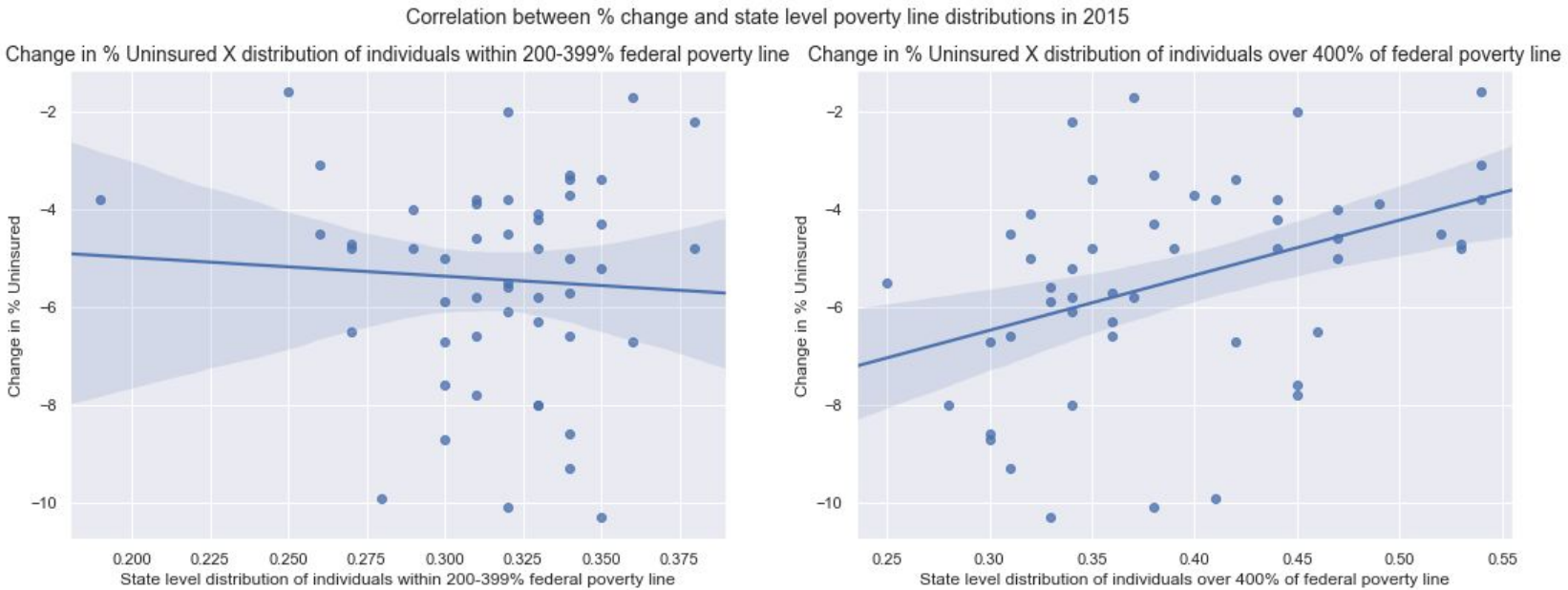
- Health Expense Per Cap 2012 (wanted 2010 data but was unavailable)
- Ratio of individuals over 200% federal poverty line (FPL) to individuals under 200% of FPL
- Uninsured Rate 2010
- Political Lean
- Medicaid Expansion Status

```
uninsured_rate_change                          1.000000
uninsured_rate_2010                            0.677093
2015_100_199_fedpl                             0.512616
health_expense_per_cap_2011                    0.498907
2015_under199_fed_pl                           0.496695
health_expense_per_cap_2009                    0.496485
2015_over200_fed_pl                            0.494866
health_expense_per_cap_2010                    0.490401
ratio_rich_to_poor                             0.488840
health_expense_per_cap_2013                    0.481453
health_expense_per_cap_2012                    0.481453
2010_100_199_fedpl                             0.467727
health_expense_per_cap_2014                    0.455196
insurance_coverage_change                      0.454530
medicaid_enrollment_change                     0.441360
2015_under_fed_pl                              0.432882
PC1                                            0.426520
2010_under_fed_pl                              0.424199
PC2                                            0.408767
2015_over400_pl                                0.404311
2010_over400_pl                                0.370733
medicaid_enrollment_2016                       0.344929
medicare_enrollment_2016                       0.308899
marketplace_health_coverage_2016               0.305778
marketplace_tax_credits_2016                   0.304350
gdppercap_2012                                 0.291288
gdppercap_2014                                 0.285779
gdppercap_2013                                 0.284884
employer_health_coverage_2015                  0.277582
gdppercap_2015                                 0.277536
gdppercap_2011                                 0.270893
medicaid_enrollment_2013                       0.267051
private_health_insurance_spending_millions     0.261374
medicaid_expansion                             0.259846
gdppercap_2010                                 0.257893
medicaidXpolitics                              0.241486
political_lean                                 0.241486
uninsured_rate_2015                            0.224251
PC3                                            0.137967
average_monthly_tax_credit_2016                0.064360
2015_200_399_fedpl                             0.062100
2010_200_399_fedpl                             0.043874
Name: uninsured_rate_change, dtype: float64
```

Exploration - effect of poverty levels:



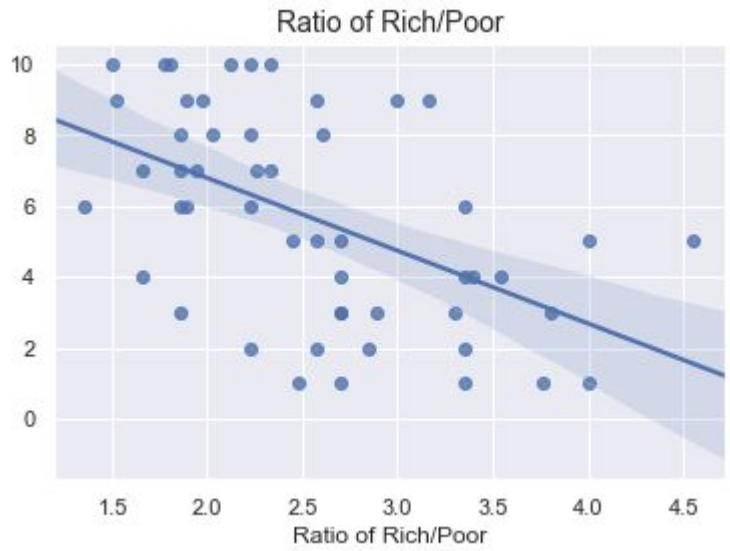
Exploration - effect of poverty levels:



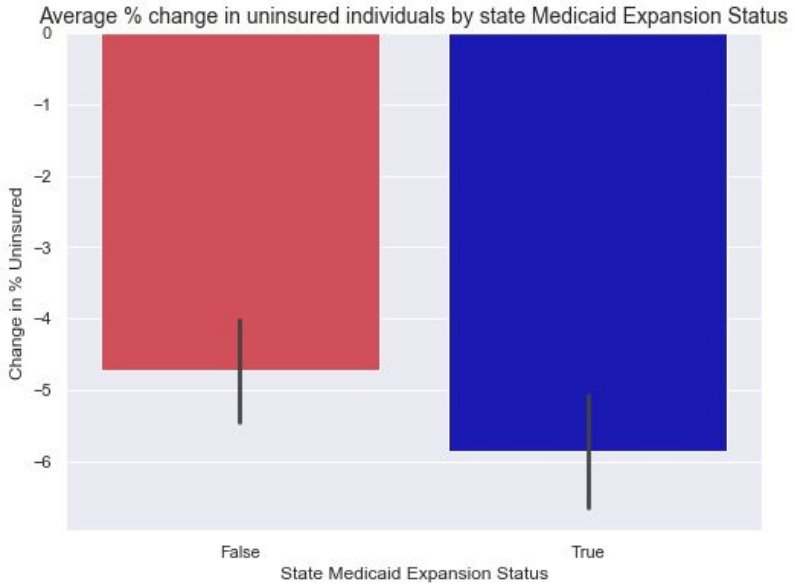
Feature ratio of ‘Rich’ to ‘Poor’ gave best balance of descriptive value and dimensionality.

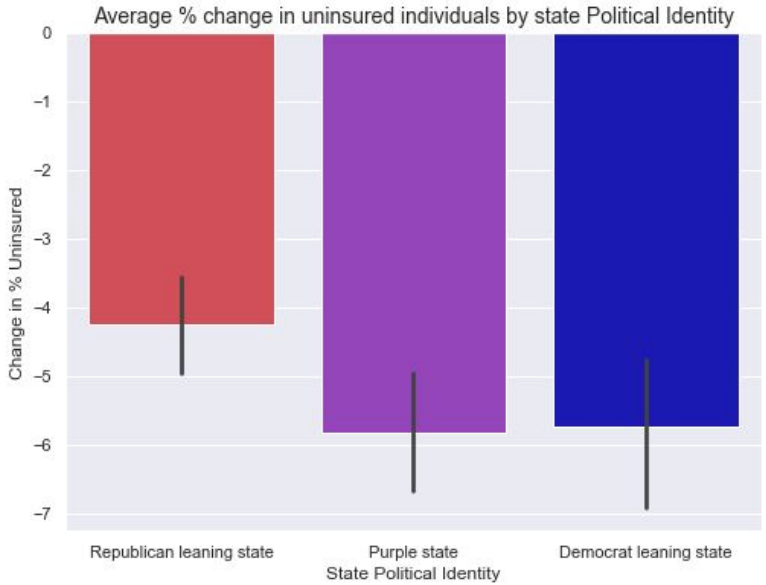
uninsured_rate_change	1.000000
uninsured_rate_2010	0.677093
2015_100_199_fedpl	0.512616
health_expense_per_cap_2011	0.498907
2015_under199_fed_pl	0.496695
health_expense_per_cap_2009	0.496485
2015_over200_fed_pl	0.494866
health_expense_per_cap_2010	0.490401
ratio_rich_to_poor	0.488840
health_expense_per_cap_2013	0.481453
health_expense_per_cap_2012	0.481453
2010_100_199_fedpl	0.467727
health_expense_per_cap_2014	0.455196
insurance_coverage_change	0.454530
medicaid_enrollment_change	0.441360
2015_under_fed_pl	0.432882
PC1	0.426520
2010_under_fed_pl	0.424199
PC2	0.408767
2015_over400_pl	0.404311
2010_over400_pl	0.370733
medicaid_enrollment_2016	0.344929
medicare_enrollment_2016	0.308899
marketplace_health_coverage_2016	0.305778
marketplace_tax_credits_2016	0.304350
gdppercap_2012	0.291288
gdppercap_2014	0.285779
gdppercap_2013	0.284884
employer_health_coverage_2015	0.277582
gdppercap_2015	0.277536
gdppercap_2011	0.270893
medicaid_enrollment_2013	0.267051
private_health_insurance_spending_millions	0.261374
medicaid_expansion	0.259846
gdppercap_2010	0.257893
medicaidXpolitics	0.241486
political_lean	0.241486
uninsured_rate_2015	0.224251
PC3	0.137967
average_monthly_tax_credit_2016	0.064360
2015_200_399_fedpl	0.062100
2010_200_399_fedpl	0.043874

Name: uninsured_rate_change, dtype: float64

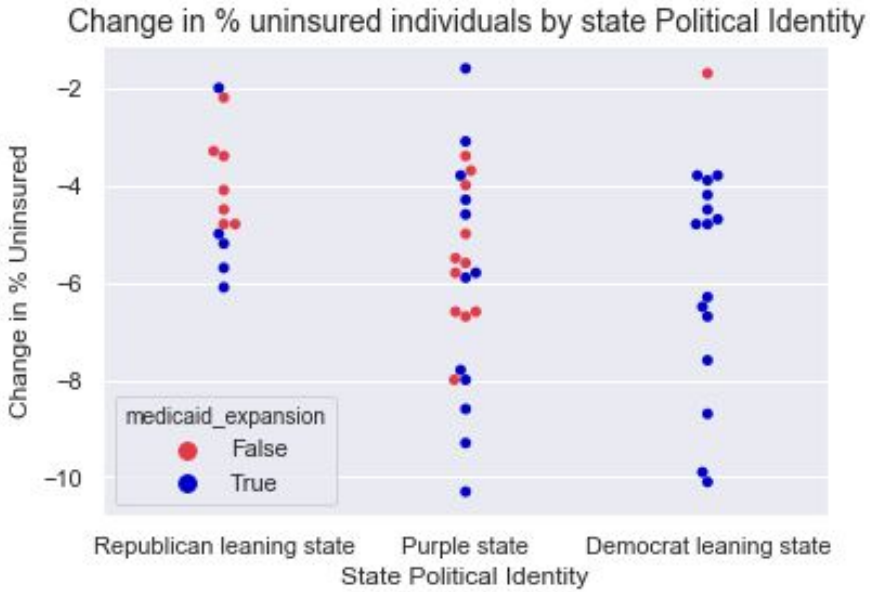


Exploration - Medicaid Expansion X State politics interaction (Pt I)

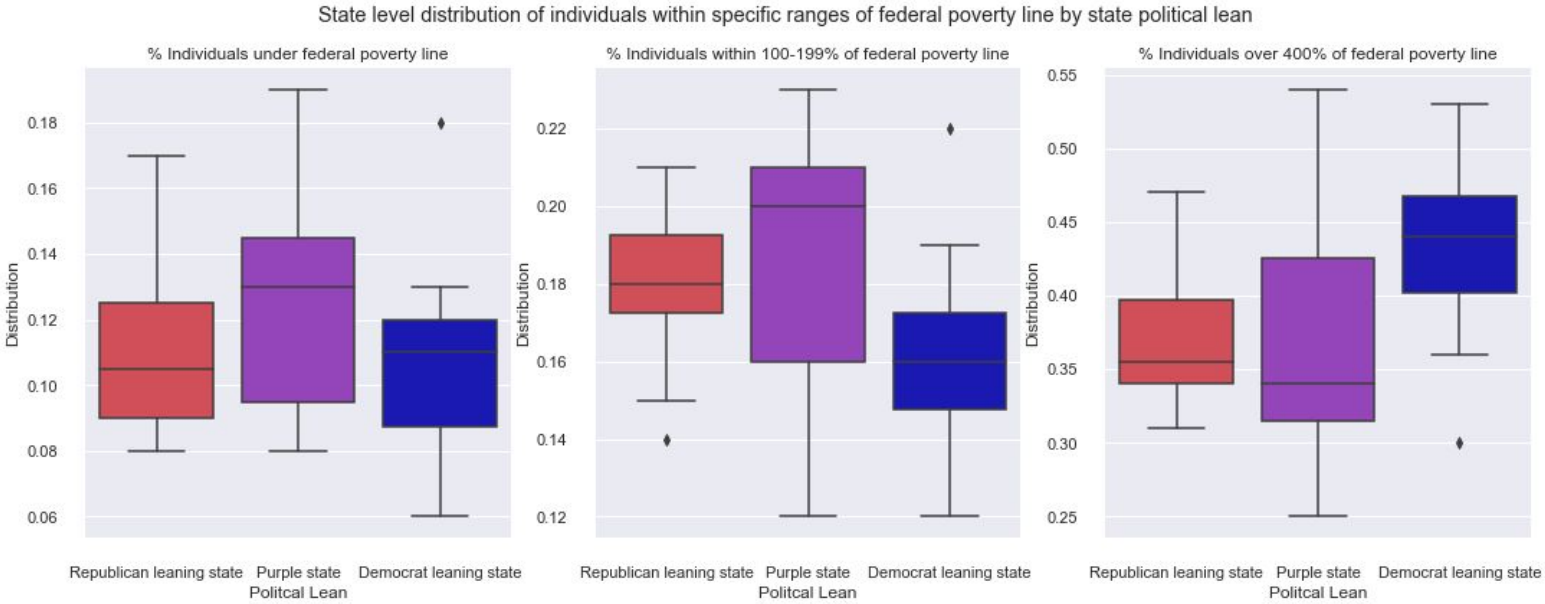




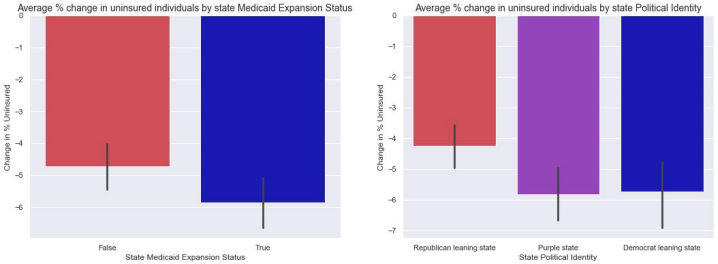
Exploration - Medicaid Expansion X State politics interaction (Pt II)

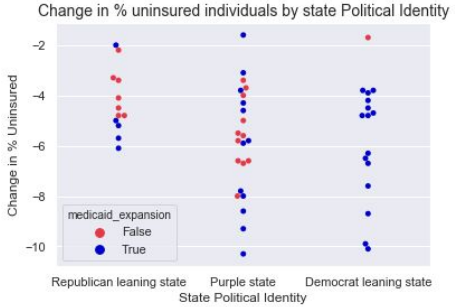


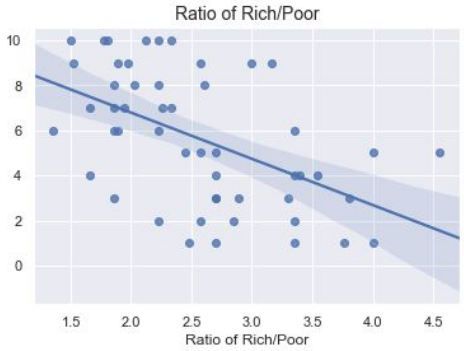
What about when we account for poverty level?

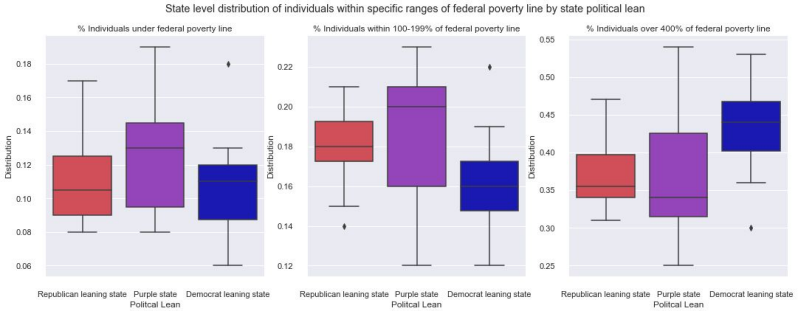


Suggests economics and medicaid expansion offer bulk of predictive power, but political lean flavors these interactions and adds extra predictive power.









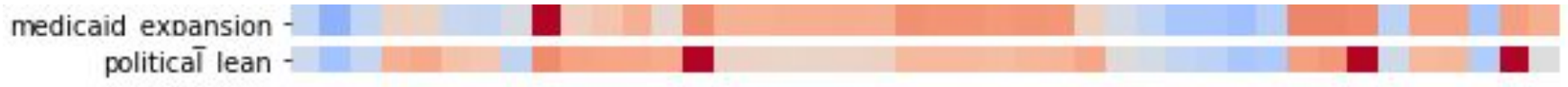
Should we include state politics as a feature variable?

This requires justification - it should only be added if it adds predictive value, not to support a partisan stand point.

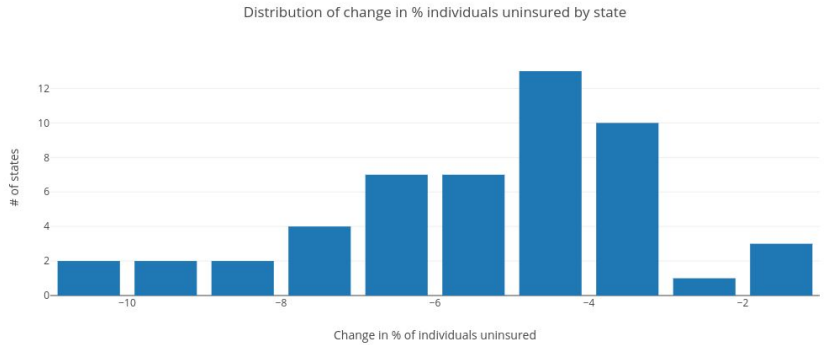
Both medicaid expansion status and political lean added significant predictivity accuracy to my models that could not be replicated with other variables.

Generally low correlation with other features of the dataset.

TLDR: Politcal lean had an effect that couldn't be replicated by inclusion of other variables.



Any issues with our outcome variable?




```
count    51.000000
mean     -5.433333
std       2.133229
min     -10.300000
25%      -6.600000
50%      -5.000000
75%      -3.950000
max      -1.600000
Name: uninsured_rate_change, dtype: float64
```

Slight negative skew.

Otherwise relatively normally distributed.

Data not transformed or modified to preserve integrity of investigation.

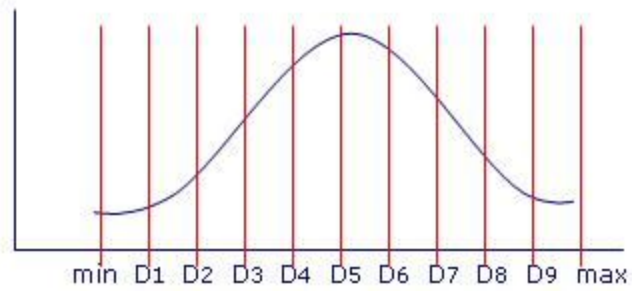
The models:

Initial approach:

- Initial models utilized all variables
- Not efficient, or productive time wise
- Subsequent models focussed on my selected variables
- Predictive accuracy increased, as to be expected
- Tried as both classification and regression problem...regression results presented here..

Some models aren't suited to regression… Can I make this a classification problem if needed?

Some supervised learning models cannot be adopted towards regression. As an option, I also encoded the data in categories depending on their decile level. Could also look at quintiles, quartiles, etc.



Downside of this approach:

- Loss of resolution could mean artificially higher score metrics for these models.
- Groups split equally in number but not in effect size bins.

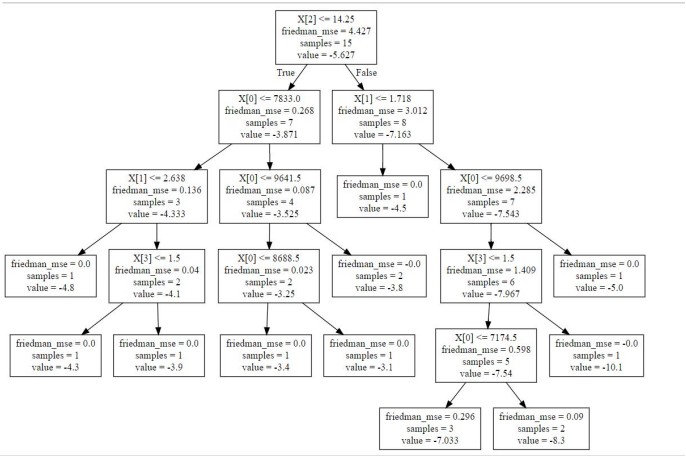
The models - Decision tree regressor:

Decision Trees:

Tried as a classification problem with deciles, not productive. Best score was 0.25

Ran as a regressor with following settings:

Criterion=**Friedman MSE [2]**
Max features=3
Max depth=5
Splitter = best



Optimizes decision tree splits based on Friedmann improvement score on MSE.

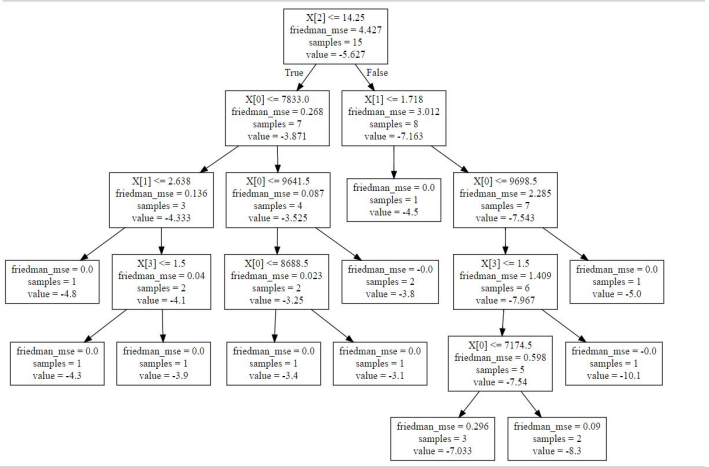
In this context, there's a problem with this.

The models - Decision tree regressor:

Decision Trees:

Regressor with following settings:

Criterion=Friedman MSE
Max features=3
Max depth=5
Splitter = best



Results:
Training R^2 = 0.999
Test R^2 = -0.114
Completely overfit

It's easy for the computer to get the MSE down and fit the data with this improvement score. We only have 52 samples.

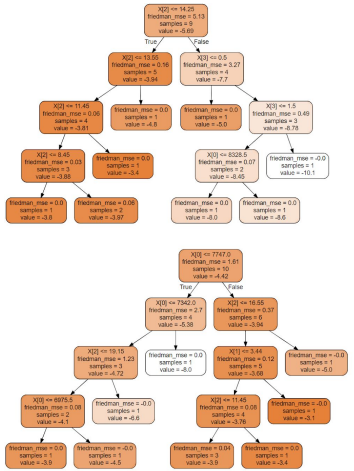
It's not a generalizable model though.

The models - Random Forest:

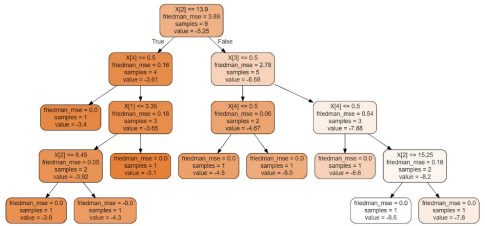
Grid Search CV yielded similar parameters:

Criterion=Friedman MSE
Max features=2
Max depth=4
# of estimators =20

Results:
Training R^2 = 0.638
Test R^2 = 0.525
Not a completely terrible model - but not still not great



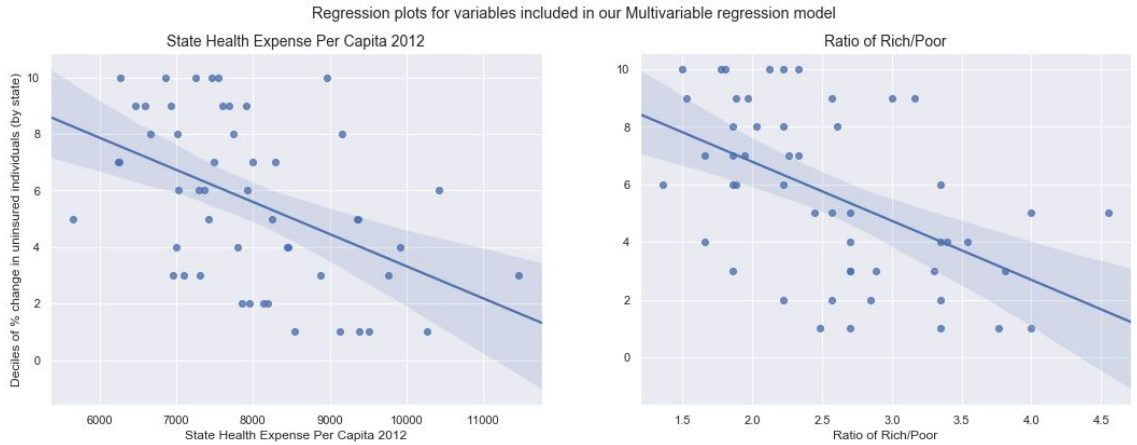
Some of the trees from my forest.



The models - Multivariable Linear Regression:

Sometimes simplest is the best:

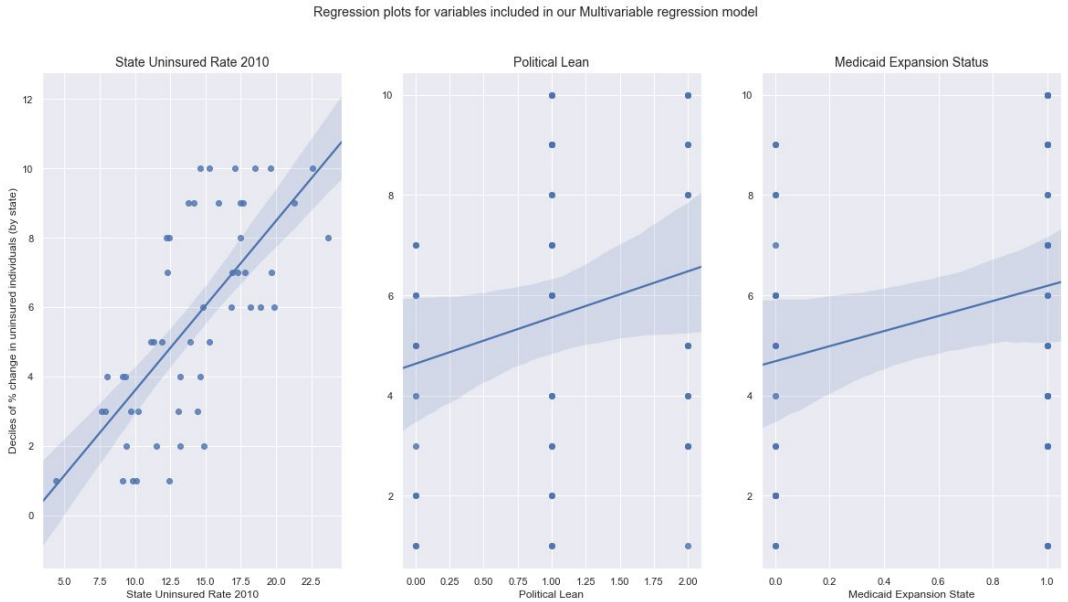
Results:
Training R^2 = 0.758
Test R^2 = 0.702
A good model



The models - Multivariable Linear Regression:

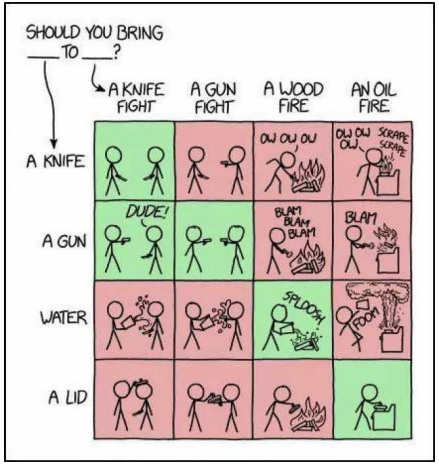
Sometimes simplest is the best:

Results:
Training $R^2 = 0.758$
Test $R^2 = 0.702$
A good model



The models - Gradient Boost Regression

Grid Search CV yielded the following parameters:
Learning rate=0.2
loss=ls
Max features=4
# of estimators =8

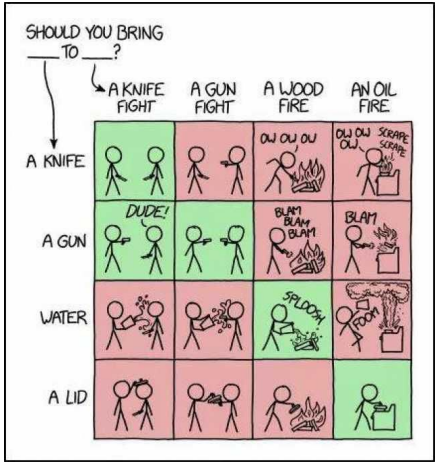


‘Bringing a gun to a knife fight’

Results:
Training R^2 = 0.674
Test R^2 = 0.30
Looks ok initially..but overfit

The models - Gradient Boost Regression

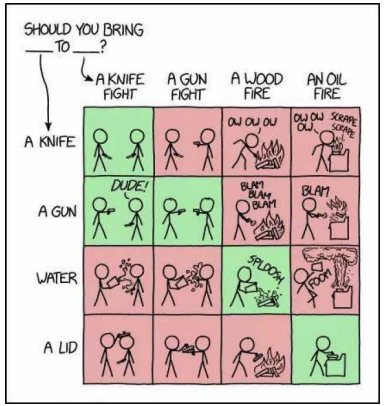
Grid Search CV yielded the following parameters:
Learning rate=0.2
loss=ls
Max features=4
# of estimators =8

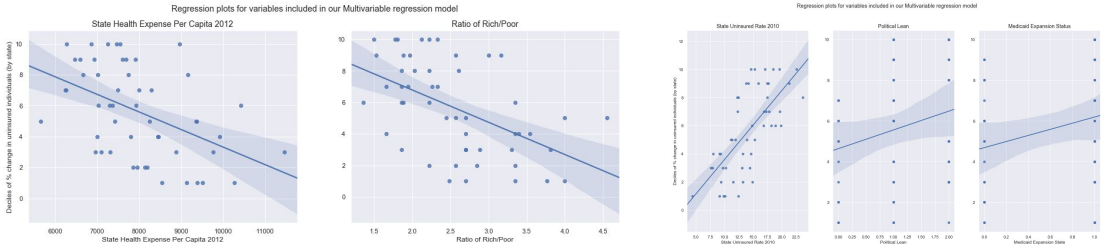


‘Bringing a gun to a knife fight’

Results:
Training R^2 = 0.674
Test R^2 = 0.30
Looks ok initially..but overfit

The models - Gradient Boost Regression





Advantages of boosting methods with linear regression:
-Helping in the case of overfitting. (helps too much here..)
-When data is of some non linear complex shape. Boosting allows model to slowly evolve to fit data.

Gradient Boost Regression results:

Training R^2 = 0.674

Test R^2 = 0.30

Not a good model

Vanilla Multivariate Linear Regression results:

Training R^2 = 0.758

Test R^2 = 0.702

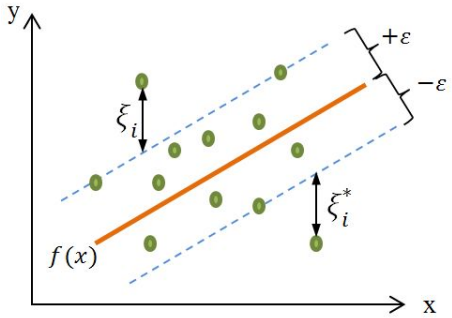
A good model

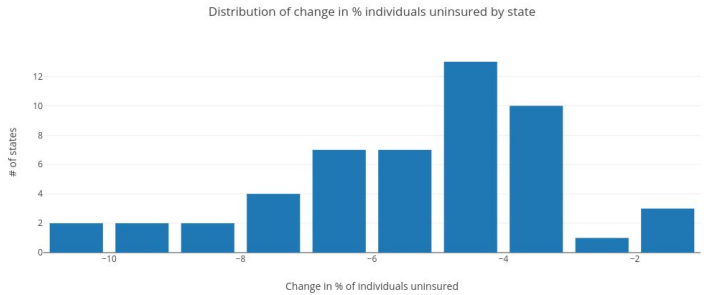
The models - Support Vector Machines

Grid Search CV crashed my kernel, so parameters were tested individually:

C: No impact from changing
Episilon: No impact from changing
Kernel: RBF or Linear worked equally (poorly)
Used RBG with default values.

Support Vector Machines results:
Training R^2 = 0.445
Test R^2 = 0.608
Underfit - model may be biased towards positive end of tail

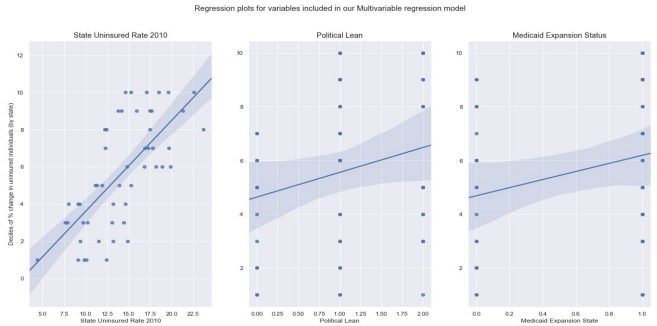




The models - my pick:

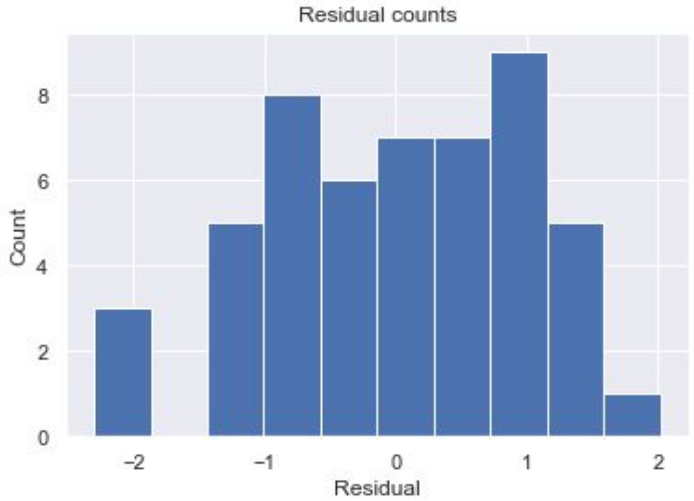
The multivariable linear regression model has high predictive accuracy AND seems to be generalizable.

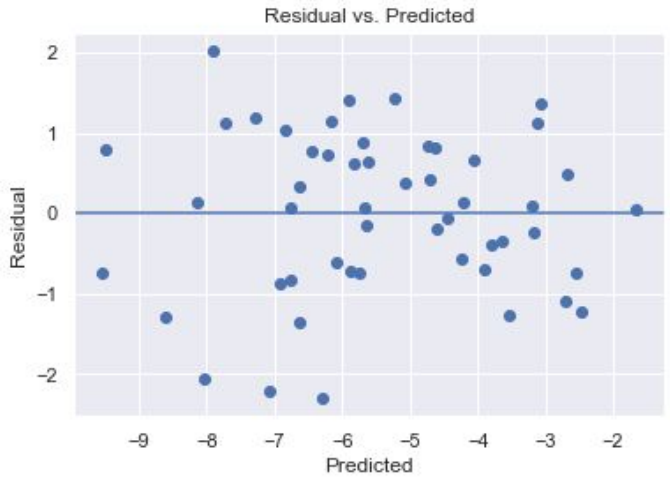
Model was not over OR underfit.



The models - my pick:

- Model prediction shows normal distribution and relative homoscedascity.
- Trend for higher residuals for states with higher changes in uninsured rate.
    - Somewhat to be expected, also may suggest some bias in model - likely due to negative skew of distribution.





What we learned...:

Real world impact (for data scientists):

- We can give an estimate on the effect any given state *could have had from the ACA under ideal conditions using our model...
- We have a better idea what variables affect the real world impact of legislation like the ACA.
- Predictive quality of analysis could be used to craft state specific legislation to reduce effect of externalities on both *economic outcomes* and *human quality of life*.
- Interested party's could use this information how they choose...

What we learned...:

Real world impact (translated):

- Change in percentage of uninsured individuals as a result of the ACA can be most accurately predicted by the following variables:
    - The baseline uninsured rate (obvious)
    - State health expense per capita (greater pressure to adopt health care coverage)
    - Ratio of individuals below 200% of FPL to above 200% of FPL (various reasons)
    - State Political Lean
    - Medicaid Expansion Status

What we learned...:

My own learning:

- Start with simpler models and then build up
- Start with fewer higher impact variables and then build up not vice versa.
- Sometimes complex problems have simple answers - you just need to know how to look for them

What's next..

To address the short comings of this model a number of steps could be taken:

1. Small sample size - 2 answers to this problem:
    a. change the way the data is sampled, maybe sample towns instead of states (downside: resource/capital intensive)
    b. Use synthetic data generation (although this is often put foward as an answer to class imbalance, it also has the potential to introduce bias..)
2. The higher residuals for those states with higher % changes - needs to be further explored. Perhaps there is potential to improve this model with further tinkering to feature selection.
3. Analysis may have some utility to individuals in public health policy - plan to simplify presentation and circulate to potential end users.

Thank you:

Dr. David Reed, My mentor

Mike Ricos, Reviewer

You, The audience

References

1. https://www.healthaffairs.org/do/10.1377/hblog20160321.054035/full/
2. https://statweb.stanford.edu/~jhf/ftp/trebst.pdf

Removing political lean....

In [18]:
```
X1=pd.DataFrame()

X1['health_expense_per_cap_2012']=state_aca_df['health_expense_per_cap_2012']
X1['ratio_rich_to_poor']=state_aca_df['ratio_rich_to_poor']
X1['uninsured_rate_2010']=state_aca_df['uninsured_rate_2010']
X1['medicaid_expansion']=state_aca_df['medicaid_expansion']

X1 = X1.dropna(axis=1)



# Instantiate and fit our model.
regr = linear_model.LinearRegression()
regr.fit(X1, Y)

# Inspect the results.
print('\nCoefficients: \n', regr.coef_)
print('\nIntercept: \n', regr.intercept_)
print('\nR-squared:')
print(regr.score(X1, Y))
```

```
Coefficients:
 [ 4.35255407e-04  5.40829099e-01 -2.79299319e-01 -2.45613369e+00]

Intercept:
 -4.796237536728434

R-squared:
0.7308521814651335
```

Based on comparison of models, effect of political lean is estimated is estimated at minimum of 7-8% of total model.

Removing Medicaid Expansion status...

In [20]:
```
X1=pd.DataFrame()

X1['health_expense_per_cap_2012']=state_aca_df['health_expense_per_cap_2012']
X1['ratio_rich_to_poor']=state_aca_df['ratio_rich_to_poor']
X1['uninsured_rate_2010']=state_aca_df['uninsured_rate_2010']
X1['political_lean']=state_aca_df['political_lean']

X1 = X1.dropna(axis=1)



# Instantiate and fit our model.
regr = linear_model.LinearRegression()
regr.fit(X1, Y)

# Inspect the results.
print('\nCoefficients: \n', regr.coef_)
print('\nIntercept: \n', regr.intercept_)
print('\nR-squared:')
print(regr.score(X1, Y))
```

```
Coefficients:
 [ 1.79822505e-04  3.75771114e-01 -3.09396530e-01 -1.17115655e+00]

Intercept:
 -2.1883659567257405

R-squared:
0.6250213759907823
```

Based on comparison of models, effect of Medicaid Expansion status is estimated is estimated at about 17% of total model

Removing Medicaid Expansion and political lean status...

In [21]:
```
X1=pd.DataFrame()

X1['health_expense_per_cap_2012']=state_aca_df['health_expense_per_cap_2012']
X1['ratio_rich_to_poor']=state_aca_df['ratio_rich_to_poor']
X1['uninsured_rate_2010']=state_aca_df['uninsured_rate_2010']
X1 = X1.dropna(axis=1)



# Instantiate and fit our model.
regr = linear_model.LinearRegression()
regr.fit(X1, Y)

# Inspect the results.
print('\nCoefficients: \n', regr.coef_)
print('\nIntercept: \n', regr.intercept_)
print('\nR-squared:')
print(regr.score(X1, Y))
```

```
Coefficients:
 [ 1.25076567e-04  2.40404622e-01 -2.93519805e-01]

Intercept:
 -2.892428391189642

R-squared:
0.46696632989470666
```

Based on comparison of models, effect of political lean and medicaid status combined is about 41% of model.

Take home -> inclusion of one of these features in the model makes up to a large degree for exclusion of another -> some covariance

Medicaid Expansion status played a bigger role in model, interpret causality and/or correlation of these effects as you may....