THE EFFECT OF CIRCUIT COURT DECISIONS AND JUDGE CHARACTERISTICS ON SOCIAL AND POLITICAL ATTITUDE OF THE PUBLIC: CAUSAL ANALYSIS

BHASKAR GURRAM
New York University
Courant Institute of Mathematical Science
bhaskar.gurram@nyu.edu

NAMAN KUMAR
New York University
Courant Institute of Mathematical Science
naman.kumar@cs.nyu.edu

PRANEETH YENUGUTALA
New York University
Courant Institute of Mathematical Science
praneeth@nyu.edu

ROHIT MUTHYALA
New York University
Courant Institute of Mathematical Science
rrm404@nyu.edu

with help from Advisors*

Abstract

We present an application of machine learning and causal inference to study the effect on social and political attitude of the public, based on court rulings and characteristics of judges involved. Our goal is to determine if the sentiment of the judges in the court rulings towards certain targets(like republicans, democrats, woman, feminists, etc) is a factor through which we can infer the public opinion towards the same targets. We not only use the judges sentiment towards these targets, but we also weight them similarities to those targets using word2vec. We also use the biographical and other characteristics of the judge(s) involved in the cases for determining this causal effect, by using them as instruments. Using them as instruments in 2-stage least squares regression helps to ensure that this determined causal effect is infact consistent and unbiased. We also find which of these characteristics are important as instruments, using different feature selection methods.

We use American National Election Survey data, Court rulings data, and judge characteristics data for this analysis into the causal effect.

*Dr. Daniel Chen and Dr. Elliot Ash

1 Introduction

There are 13 appellate courts that sit below the U.S. Supreme Court, and they are called the U.S. Courts of Appeals. The 94 federal judicial districts are organized into 12 regional circuits, each of which has a court of appeals. The appellate courts task is to determine whether the law was applied correctly in the trial court. Appeals courts consist of three randomly selected judges and do not use a jury. Each judge has a different background and characteristics.

The rulings of these courts can be a major factor in shaping the social and political attitude of the public. This is the basic premise that we base this project on. The rulings themselves are verbose written text of several paragraphs. The social attitude of people in a certain year towards certain targets(like republicans, democrats, woman, feminists, etc) are obtained through American National Survey data. We calculate paragraph and case level similarities using word2vec to these targets. We also use NLP to calculate the judge level sentiment for each case that could be used to understand the context of these paragraphs. These sentiments are aggregated at a case level first and then at the circuit-year level, by getting the weighted average of the sentiment, with similarities of the cases being the weight. Finally, we find the effect of this weighted sentiment towards the attitudes of the public.

2 Datasets

We used three datasets.

2.1 Court Rulings/Opinions

This dataset is the case level opinion of every case collected yearly from 1880 to 2013. Each case file contains a list of paragraphs of the case proceedings.

2.2 BloombergCASELEVEL_TOUSE

This database consists of case level details of each case and contains 387,898 case records from 1880 to 2013. Each case record has 134 features which corresponds to general information for the case.

2.3 BloombergVOTELEVEL_TOUSE

This dataset consists of case level details including judges biographical characteristics. It contains 1,163,694 vote records and has 414 features. For each case, there are 3 records; one for each judge, with the common case characteristics and individual judge characteristic as features. The BloombergVOTELEVEL_TOUSE contains all the information contained in BloombergCASELEVEL_TOUSE and more. Hence, we mainly focused on the vote level data set.

2.4 American National Election Survey

This dataset consists of the results of surveys conducted between 1948 and 2008. Each record in these surveys corresponds to the answers of each individual participant. The questions that has been asked in three or more Time Series studies (once every two years) are included in this data. There are 951 distinct questions that were asked across the years, 40 of these are chosen by us, known as thermometer variables(or targets, which we alluded to in the introduction section).

The value of assigned to a thermometer variable/target for an individual is decided based on the following:

1. A value around 50 is assigned to a thermometer target for an individual, if the individual lacks opinion/knowledge about the target (group) in question.

2. A value between 50 to 100 is assigned to a thermometer target for an individual, if the individual feels favourably towards the target (group) in question.

3. A value less than 50 is assigned to a thermometer target for an individual, if the individual disapproves of the target (group) in the question.

3 Data Processing

All the data was accumulated at Circuit year level for causal inference. Following variables were computed/aggregated for each circuit year to be used for result computation.

3.1 The Endogenous variable - in our causal equation

From the ANES(American National Survey) dataset we obtained the thermometer variables/targets against which each case opinion had to be processed. Our endogeneous variable is the judge level sentiment towards a target, weighted by similarities to that target.

3.1.1 Data Cleaning

For sentiment and similarity computation, cases with zero paragraphs were identified and removed. For case level aggregation, cases with similar id and content were filtered. For judge level aggregation, judges with no caseids matching the cases in the sentiment and similarity corpus were dropped. We also identified targets about which the individuals were not surveyed on in a particular year/ or have invalid NaN values in the dataset.

3.1.2 Paragraph Similarity to thermometer variable

The similarity score was the cosine similarity of each paragraph to each thermometer target. The score was computed using word2vec.

3.1.3 Paragraph Sentiment

The sentiment score for each paragraph was either positive(+1), negative(-1) or neutral(0). The score was computed using the open-source Vader sentiment analyzer.

3.1.4 Computation

The data was from a case i with panel of judges J, where judges were indexed by j, in circuit c, at year t. There are a set of thermometer targets, indexed by k. Each paragraph d in case i has a similarity to each target, W_{id}^k, and a sentiment, Sid. Let W_i^k denote the vector of paragraph similarities in case i. Let S_i denote the vector of sentiments in case i. The case-level sentiment towards target k is defined as: $S_i^k = W_i^k S_i$, which is the dot product. The circuit-year level endogenous regressor is therefore S_{ckt}, the average case-level sentiment toward k for each case in circuit-year ct.

3.2 Instrument variable

For the causal inference test we created our initial set of instrument variable by processing the BloombergVOTELEVEL_TOUSE dataset

3.2.1 Data Cleaning

The ANES data had many missing thermometer target values. While aggregating over a particular circuit year for a particular thermometer, we only took those respondents values which have been answered for that thermometer target.

3.2.2 Weighted average judge biographical characteristic

Z_{ckt} - weighted average judge biographical characteristics in circuit c and year t, weighted by the similarity to target k of the cases to which the judges are assigned.

3.2.3 Variables to remove endogeneity bias

1. γ_{ck} - a set of dummy variables(fixed effects) for each circuit-target
2. γ_{kt} - a set of dummy variables(fixed effects) for each year-target
3. γ_{ct} - a set of dummy variables(fixed effects) for each circuit-year

3.3 Exogenous variable/ Outcome variable - Y for 2nd stage regression:

The outcome variable is defined by circuit c and year t and target k. It is the average thermometer score Y_{ckt} for all respondents in the ANES in that circuit-year.

We cleaned and processed the ANES dataset to get the average feeling thermometer value for each thermometer. The feeling thermometer value was averaged on a circuit-year level. There are 12 circuits in USA which is divided based on state. Multiple states goes to a single circuit. The mapping of state code to circuit number was used to aggregate the survey records on the circuit-year level. The generated matrix of dimension - number of circuit years * number of feeling thermometer variables is flattened across columns to get the Y_{ckt} vector.

4 Modeling

4.1 Word2vec

4.1.1 Definition

Word2vec are a group of models that are used for word embeddings,[2] they are used to represent words in vector spaces. Contextually similar words have similar embeddings and so cosine similarity can be used as a measure to get the similarity of words.

4.1.2 Training

The individual case level data was cleaned and processed to train a word2vec model. Our model represents each word by a vector of length 100. The opinion/rulings dataset was used for the training of word2vec.

4.2 Similarity Computation

We used a vectorized approach to optimize the performance and reduce computation time. A vector of size 100 was generated for each word using the word2vec model and the average vector is computed using the vectors of the words in those paragraphs. Cosine similarity was used to compute the similarity of each paragraph to the target thermometer variables.

4.3 Sentiment Computation

Sentiment analyzer of Vader from python NLTK was used to get an aggregate sentiment score for each paragraph in all the cases.

4.4 Feature Selection

Feature selection is a method in Machine Learning in which we select a subset of relevant features which can be used in model construction[7]. It is also known as variable selection or subset selection. Feature selection techniques are used for four reasons:

1. The models can be simplified and hence it makes much useful to be interpreted for researchers

2. The training time becomes much shorter

3. Dimensionality is reduced

4. Overfitting is reduced(reduction of variance)

We used ElasticNet and Lasso and cross validation with them to find the instruments which lead to a higher first stage F-statistic value in the first stage regression.

4.4.1 LassoCV

Lasso is a linear model with iterative fitting along a regularization path[8]. The best model is selected by cross validation. The optimization objective for Lasso is: $\frac{||y-Xw||_2^2}{2*n_{samples}} + \alpha||w||_1$

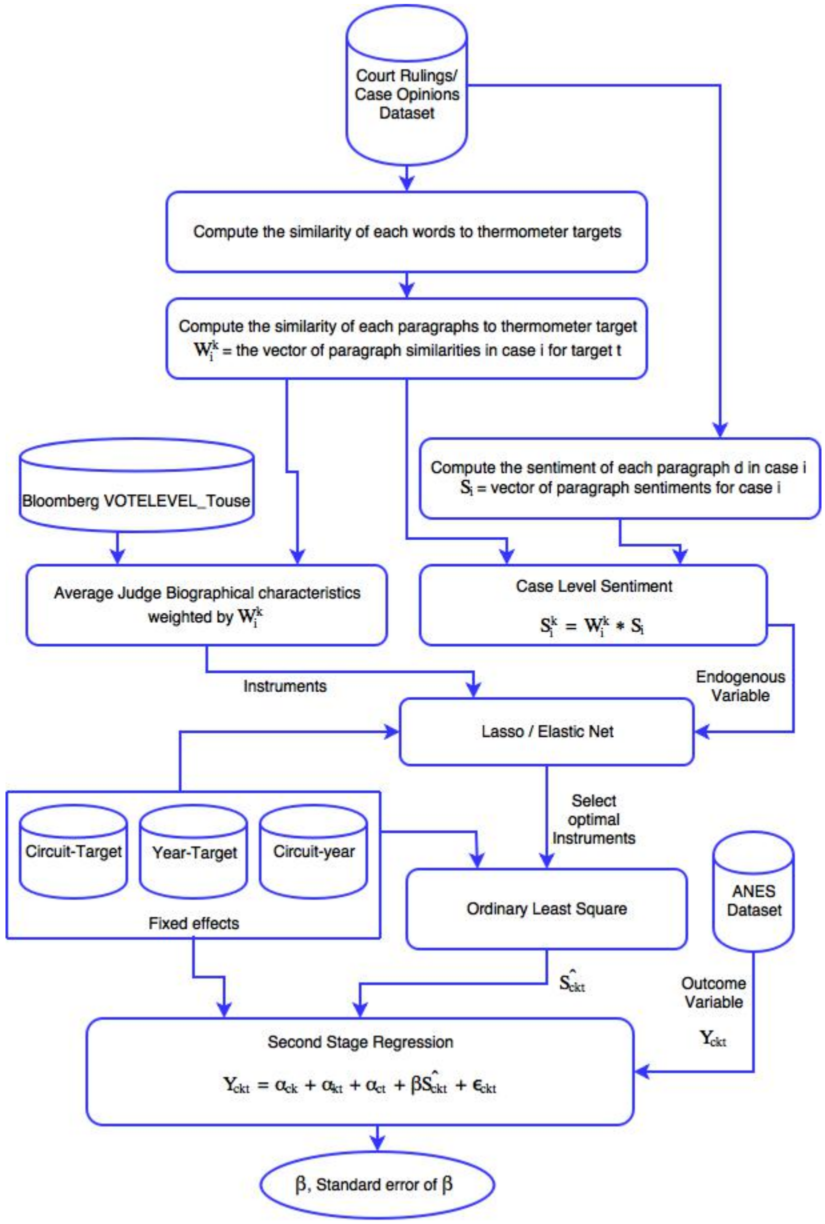
4.4.2 ElasticNet

It is a regularized regression method in which it combines both Lasso and Ridge regression methods[9]. The optimization objective is given as $\frac{||y-Xw||_2^2}{2*n_{samples}} + \alpha * l1_{ratio} * ||w||_1 + 0.5 * \alpha * (1 - l1_{ratio}) * ||w||_2^2$

4.5 Ordinary Least Square Regression

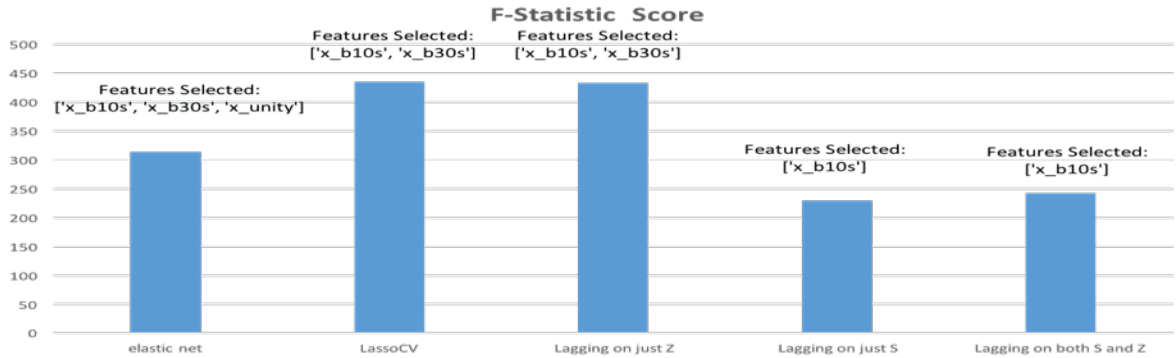
It is a method of computing the unknown parameters in a linear regression model[10]. This method minimizes the sum of the square of the difference of the actual value and the observed value. We have used Ordinary Least Squares in the first stage primarily as a means to remove endogeneity in our endogeneous variable for the second stage regression, and also used it as a means to find out the instruments which lead to a higher F-stat value in the fist stage.

5 Design Diagram



The design diagram as shown above depicts the overall workflow of the project.

6 Results



While performing the second stage regression, we have used LassoCV as the feature extraction method.

Table 1: Second Stage Results

method	coef	std err	t	$P > \|t\|$	0.025	0.975
No lag	-143.0589	44.051	-3.248	0.001	-229.420	-56.698
Lag on both S and Z	-257.4710	45.194	-5.697	0.000	-346.074	-168.868

Table 2: Other Experiments Second Stage Results

method	coef	std err	t	$P > \|t\|$	0.025	0.975
Lag on S	-130.0840	42.898	-3.032	0.002	-214.185	-45.983
Lag on Z	-257.4727	45.193	-5.697	0.000	-346.075	-168.871

Top words similar to target variables

Thermometer Variable	TOP-1	TOP-2	TOP-3	Thermometer Variable	Top-1	Top-2	Top-3
Democrats	Committeemen	Electors	Votes	Political Independents	Political	Partisan	Ideological
Republicans	Voters	Candidates	Canvassers	Political Parties	Parties	Litigants	Tribes
Protestants	Intervenors	Respondents	Petitioners	Poor People	Poor	Gentle	Habits
Catholics	Muslims	Indonesians	Coptic	Republican Party	Party	Nonparty	Creditor
Jews	Christians	Russians	Germans	Women Right Activist	Right	Desire	Opportunities
Blacks	Negroes	Minorities	Nonwhites	Young People	People	Teenagers	Young
Whites	Caucasians	Anglos	Promotions	Asian Americans	Americans	Asian	African
Southerners	Villains	Traitors	Ninja	Congress	Legislature	Legislation	Drafters
Big Business	Business	Big	Thriving	Environmentalists	Longliners	Fishers	fishermen
Labor Unions	Labor	Unions	Union	Anti-Abortionists	Anti	Combats	Prevention
Liberals	Gruffalo	Muggers	Baruch	Federal Government	Government	Federal	State
Conservatives	Greeks	Italians	Partisans	Illegal Aliens	Aliens	Illegal	Immigrants
Military	Civilian	Naval	Uniformed	Christian Fundamentalists	Christian	Pentecostal	Buddhist
Policemen	Patrolmen	Youths	Officers	Radical Students	Students	Pupils	Schools
Black Militants	Black	White	Hispanic	Farmers	Growers	Dairies	Merchants
Civil Rights Leaders	Rights	Civil	Interests	Feminists	Entrepreneurs	Ambassadors	Competitors
Chicanos Hispanics	Hispanics	Latinos	Chicanos	Evangelical Groups	Groups	Churches	Congregations
Democratic Party	Litigant	Candidate	party's	Elderly	Infant	Homeless	Autistic
Middle Class People	Residents	Class	counties	Supreme Court	Supreme	Court	Sjc
People on Welfare	on	upon	lives	Women	Females	Males	youngsters

7 Performance Evaluation

7.1 Feature Selection

We used Lasso, elastic-net methods for feature selection. Features selected via Lasso give the best F-stat value. The features selected are "x_b10" and "x_b30". "x_b10" represents if the judge is born in 1910's and "x_b30" represents if the judge is born in 1930's.

7.2 F-TEST

The test is performed after first stage regression. A high value of F-stat (higher than 12) indicates a strong first stage.

7.3 Using Lag in Various Settings (Advanced Feature Selection)

Lag values for Endogenous and Instrument variables is used. Results are computed using multiple such values for best setting.

7.4 Endogenity Bias Handling (Fixed Effect Modeling)

Dummy variable is plugged in to account for fixed affect. Missing variable bias/ endogeneity bias is hence accounted and removed.

7.5 Confidence Interval

Confidence interval for second stage regression is determined. High absolute value in the confidence interval signals more confidence on conclusion.

7.6 P-Value

Small p value signifies a good prediction. A p value below 0.05 is considered good, which is achieved.

7.7 High Value Of Coefficient

The coefficient of second stage regression is generally used to evaluate result. A high absolute value signifies a good result. A high absolute value is achieved in our experiments.

8 Challenges

1. Memory and number of file limitation on NYU HPC Prince cluster was of hindrance.
2. Common features being selected through various feature selection settings limited the results and hence conclusion.

3. Due to large data size, it couldnt be loaded in memory and hence had to be split in chunks.

4. Word2vec had to be enhanced to Phrase2vec to account for phrases as thermometer variables.

9 Future Work

There are many different algorithms and topics to explore in feature selection and causality.

1. It would be interesting to analyze the estimation of heterogeneous effects using random forests developed by Athey-Wager.

2. We can implement lagged effects using the weighted features from many previous years, instead of just one year to factor in its effects and see the change in our result and conclusion.

3. Additionally, we could perform robust casual inference if we can get the public opinion data towards the thermometer variables for every year from 1880 to 2012. The constraint on availability of public opinion data restricts our research.

10 Acknowledgement

We would like to thank Dr. Daniel L. Chen and Dr. Eliiot Ash for providing the data source and guidance on the project.
We would like to extend our gratitude to Prof. David Rosenberg and Brett Bernstein for their explanations of Machine Learning principles and concepts in the course.

References

[1] Chen, Daniel L., Vardges Levonyan, and Susan Yeh. Do policies affect preferences? evidence from random variation in abortion jurisprudence. Working paper, ETH Zurich, 2014.

[2] Ash, Elliott, and Daniel L. Chen. "Religious Freedoms, Church-State Separation, and Religiosity: Evidence from Randomly Assigned Judges." (2016).

[3] "Court Role and Structure." United States Courts. N.p., n.d. Web. 14 May 2017. http://www.uscourts.gov/about-federal-courts/court-role-and-structure

[4] "Empirical Methods" http://web.mit.edu/14.771/www/emp_handout.pdf

[5] "Fixed Effect Models" http://www.jblumenstock.com/files/courses/econ174/FEModels.pdf

[6] "Word2vec." Wikipedia. Wikimedia Foundation, 11 May 2017. Web. 14 May 2017. https://en.wikipedia.org/wiki/Word2vec

[7] "Feature Selection." Wikipedia. Wikimedia Foundation, 12 May 2017. Web. 14 May 2017. https://en.wikipedia.org/wiki/Feature_selection

[8] "3.2.4.1.3. Sklearn.linear_model.LassoCV." 3.2.4.1.3. Sklearn.linear_model.LassoCV Scikit-learn 0.18.1 Documentation. N.p., n.d. Web. 14 May 2017. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html

[9] "Sklearn.linear_model.ElasticNet." Sklearn.linear_model.ElasticNet Scikit-learn 0.18.1 Documentation. N.p., n.d. Web. 14 May 2017. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

[10] "Ordinary Least Squares." Wikipedia. Wikimedia Foundation, 21 Apr. 2017. Web. 14 May 2017. https://en.wikipedia.org/wiki/Ordinary_least_squares