Ideology and Identity; Machine Learning Application

Colin Wick

5/11/2021

Abstract

In the political science and public opinion literature, the standard OLS method for estimating and predicting effects is in conflict with the highly interconnected nature of identity and ideology. In this paper I attempt two methods for incorporating these methods; regression tree and step regression. Though the success of these methods are small, they each yield results not picked up by OLS. Both have increased predictive power over OLS and yield variable combinations which could not have been constructed a priori.

Introduction

Identity and political ideology are usually examined in a linear and additive analytic structure. In polling, crosstabs are broken out by age, race, and education as discrete variables for example. In Napier & Jost (2009), a model is put together which stratifies outcome variables based on a matrix of correlations between qualities, which accounts for the significant heterogeneity and interaction between identity variables.

In this paper I present two basic models for incorporating highly-interacted models of public opinion using the ANES time-series dataset by building an aggregate ideology variable from 35 discrete attitudes and attempt to build predictive models around this ideology structure to show the potential for increased power by allowing for more freedom in variable interaction.

The general setup of the model will operate using a few rough cuts for simplicity and analytic viability. To define terms:

1. Voter demographics (DEMO) - components of a voter's identity which are "fixed" in the sense that they are reasonably exogenous to political perspective. Attitudes might plausibly flow from these characteristics, but politics would not reasonably influence these attributes.
2. Voter inputs (INPUT) - somewhat exogenous events or characteristics which may influence a voter's attitude, like being contacted by a campaign, job status (union, laid off, industry, etc), urban/rural residence, or any sort of nudges which a voter may encounter.
3. Politicization (POL) - Characteristics which are endogenous to a degree. Events known to have a radicalizing effect, political knowledge, other variables of interest along similar lines.

(Note: Due to computational complexity and data availability, this list was reduced significantly but in a more formal research setting it could be expanded, imputed, or accomodated)

The theoretical basis of the model is that there are characteristics of people's demographic or exogenous experience which influence their relative radicalism in the political system. Political science and sociological literature explores material and social condition as a generator of political opinion.

Methods

Data

The ANES is a gold-standard representative sample of the American electorate, but it faces limitations for the scope of this study. Namely, measuring actual *radical* sentiment must be inferred based on self-reported actions and attitudes, meaning some voters will be included or excluded from the classification based on the scope of questions which may not capture their underlying attitude.

The ANES contains about 1800 variables, so a pruning and variable classification process has to take place before any analysis can even be done. Ultimately, there is a degree to which variable selection is *arbitrary*. Working with public opinion data requires a degree of judgement with respect to variable selection and engineering. The ANES contains 1800 variables which overlap on some matters and contain redundancies. Such a large pool of variables makes variable construction a combinatorically unfeasable.

The data featured a significant number of *NA*s of varying random distribution. This limited both the time-series element and inter-year avenues for analysis. The most glaring issue along these lines was the changes ANES makes after every survey wave, especially in the past 20 years, which render large groups of variables unusable for regression analysis.

Similarly, within-year “Refuse” or “Inappropriate” responses make an entire observation unusable in a regression. Variables which feature a significant number of “Inappropriate” were thrown out. This affected both the ideology variable construction but also the number of “demographic” or pseudo-demographic variables available for these models.

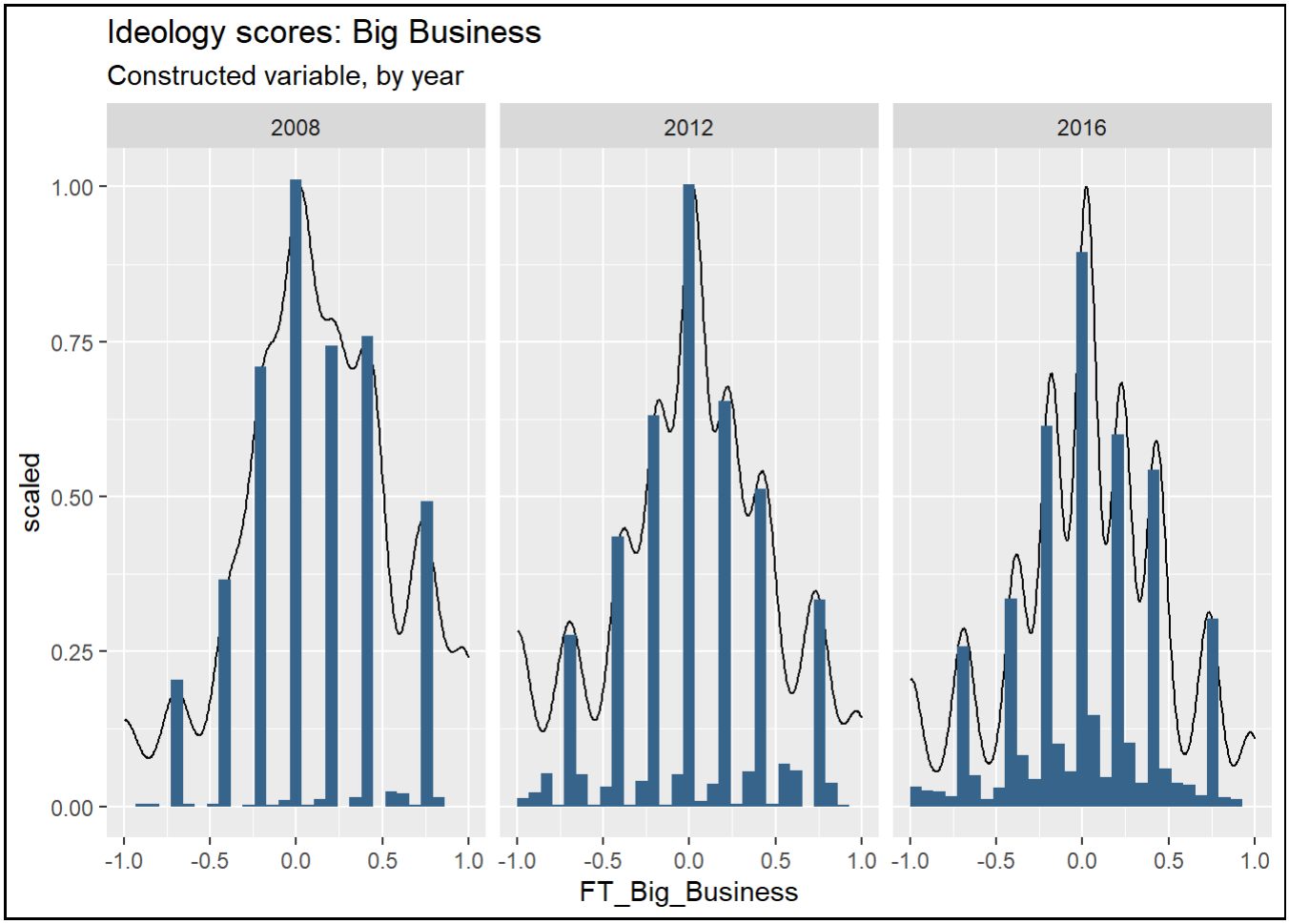
Theory and Variable Construction

Ideology itself is a slippery measurement, since political scientists and pollsters cannot simply ask “Are you ideological?” to a respondent and reliably get a usable response. An alternative is to estimate a a rough composite of ANES feelings thermometer and survey questions about non-voting civic behavior, like protesting, discussion, and activism. Variables explicitly addressing attitudes for particular candidates and positions were omitted. Instead, this study focused on questions within general public opinion at the time and voter characteristics.

The general principle of ideological positioning follows Tomz and Van Houweling (2008) and Luskin (2005) which uses self-reported stances on issues as a baseline. There is an obvious issue with taking self-reported positioning as the “true” population-wide position, but we can disregard this issue for two reasons. First, we can rest on an assumption that a voter’s error in estimating their position is, to some degree, randomly distributed. For any given voter, they may estimate themselves to be particularly liberal or moderate on an issue and report as such, but this will be relative to their network and experiences (Sinclair, 2012). We can assume some portion of variance in the difference between reported and “true” positioning. Second, this report focuses on self-conscious ideology, so a particularly liberal person in a conservative region or network ought to be picked up as “liberal” even if their “true” postion would be considered moderate by someone on the left.

There is a rough measure of ideology built-in to the ANES with the “left right self-placement” variable, which provides a baseline estimate of a voter’s self-conscious ideology. I scale each variable against its total scale, since some are on 100-point, others 7-point, or even 3-point scales by subtracting then dividing by the midpoint. For this measure, variables which would rank as “high” for liberal ideology are multiplied by -1 to put all left opinions as less than 0 and all right opinions as greater than 0.

$$V_{j,i} \in [-1, 1] = \frac{x_i - s}{s} \cdot -1^{I(lib)}$$



The visual above shows a histogram of ideological attitude towards “Big Business” for three ANES years.

For strong ideology, the Feelings Thermometer and Political Attitude measures are centered on a [-1,1] scale from left to right and summary statistics are taken. The average determines ideological position aggregated across issues, and standard deviation of this measure indicates ideological coherence of these positions. Low standard deviation respondents would correspond to those with more coherent ideological structure.

Each of these scores is summed up to an ideology score, where respondent i with consistently similar ideological positions on an issue would end up with relatively extreme positions, while those with moderate positions or cross-cutting positions end up on the tails of the distribution.

$$Ideo_i = \frac{\sum_1^j V_{j,i} - \mu_V}{\sigma_V}$$

I then normalize this measure to provide some sense of “unit”. Those with a score of 0 would be considered the average voter, swing voter, or politically unsophisticated. A score greater than 1 would represent the upper 1/3 of the population of right-wing ideological consistency and less than -1 liberal by the same definition. This provides a workable, relatively continuous variable which can be used both for classification (i.e. “The 20% most ideological voters”) and as an measure (i.e. “Participation in X is associated with 10% higher ideological consistency”).

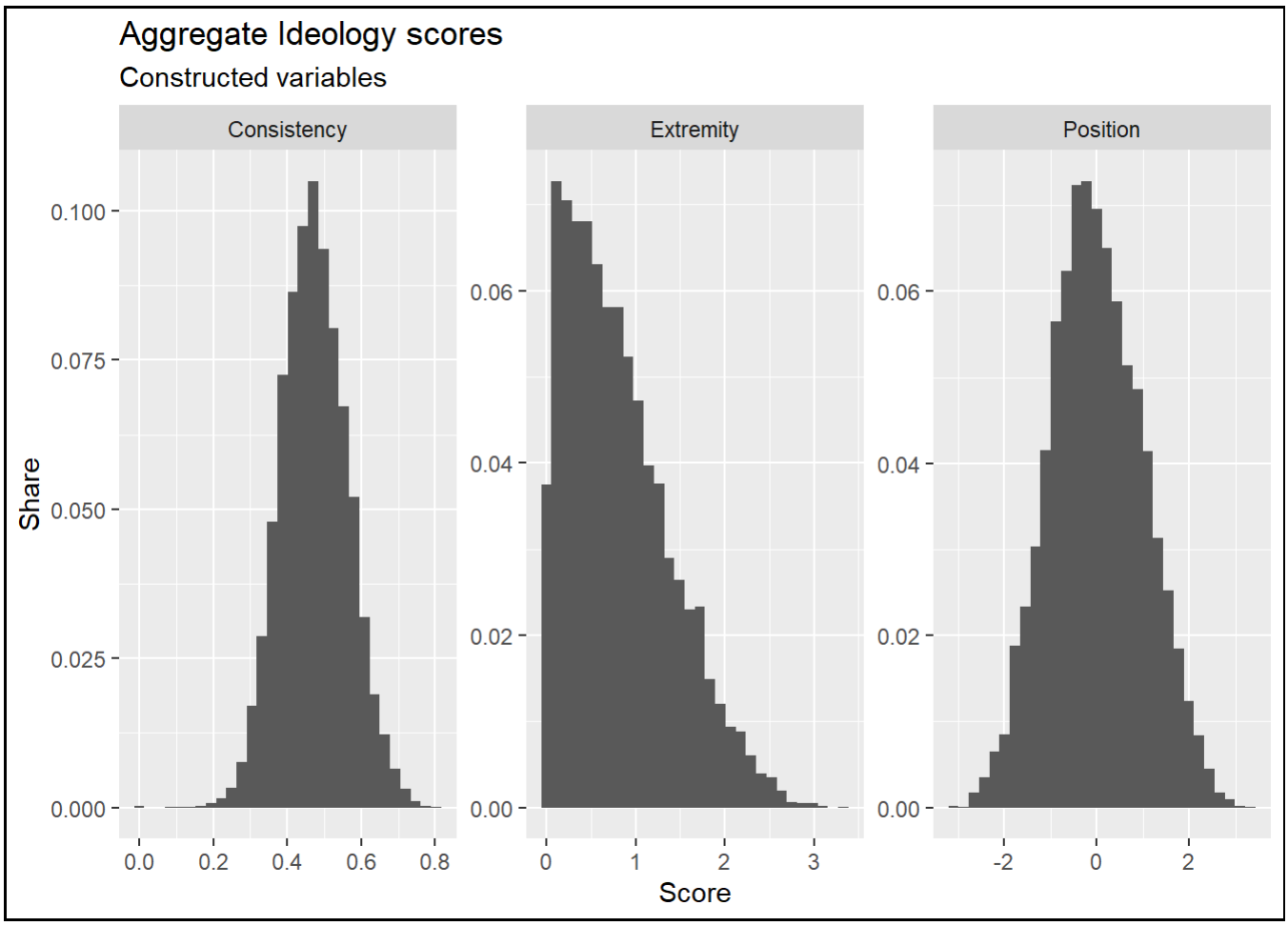
Ideological extremity is the absolute value of this measurement. This groups all those with any extreme ideologies against the same scale. Creating this variable helps distill potential variables out which may “equalize” in their predictive value if they are consistently present among those on either side of the ideological spectrum.

$$Ext_i = abs(Ideo_i)$$

Finally, ideological consistency is a measure of the average distance (standard deviation) from the aggregate ideology score and each belief held by a voter. Those with a low score will have responded with roughly the same values for each ideology question, those with a large “consistency” score will be less reliable.

$$Cons_i = \sqrt{\sum (V_{i,j} - \bar{V}_i)^2}$$

This creates the following distributions in the data.



A key issue in this dataset, and with data scientific methods in general, is rendering a causal relationship. The data used in this study is unlinked over time, which already presents issues with causal inference.

Rather, this study focuses on unearthing potentially valuable interaction terms.

Models

This study uses three model types to predict ideology from demographic characteristics at different levels of simplicity to demonstrate the effectiveness of machine learning models in dealing with public opinion study.

1. OLS - the gold-standard method for generating academically useable results. This method generates clean estimates of coefficients but lacks the hierarchical and highly-interacted structure characteristic of ideology formation as highlighted in the literature.
2. Regression Trees - Inherently accommodate hierarchy and interactions, but lack clear coefficients and statistical validation measures.
3. Step regressions - Can incorporate interactions like the tree regression, but are suceptible to misspecification of models.

Results

OLS

As a baseline for model effectiveness, I run a standard linear regression on the three parameters of interest. This first set of regressions linearly and additively regresses each ideology component on a variety of demographic and attitude variables.

```
##
## Regression Results
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                                  Position          Extremity         Consistency
##                                     (1)               (2)               (3)
## --------------------------------------------------------------------------------
## age                           -0.004 (0.004)    -0.001 (0.002)     0.001*** (0.0004)
## agesq                        0.00003 (0.00004) 0.00001 (0.00002) -0.00001*** (0.00000)
## hisp                         -0.080** (0.036)    0.001 (0.021)      -0.002 (0.003)
## own_home                      -0.008 (0.028)   -0.033** (0.016)     -0.003 (0.002)
## married                      0.084*** (0.026)   -0.001 (0.015)      0.005** (0.002)
## class_conscious               -0.009 (0.026)    -0.024 (0.015)      -0.001 (0.002)
## discuss                        0.006 (0.029)   0.076*** (0.017)     -0.002 (0.003)
## money_stocks                  -0.026 (0.028)    -0.002 (0.016)      -0.003 (0.002)
## college                       0.050** (0.025)    0.012 (0.015)       0.003 (0.002)
## black                          0.044 (0.037)     0.006 (0.021)      -0.002 (0.003)
## female                        -0.025 (0.023)    -0.019 (0.013)      -0.003 (0.002)
## protestant                     0.062 (0.039)     0.019 (0.022)      -0.003 (0.003)
## catholic                      0.101** (0.042)    0.004 (0.024)      -0.001 (0.004)
## inc_firstthird               0.086*** (0.030)   -0.010 (0.017)       0.001 (0.003)
## inc_upperthird                 0.013 (0.031)    -0.026 (0.018)       0.003 (0.003)
## lgbt                          -0.074 (0.065)     0.015 (0.037)     -0.015** (0.006)
## middleclass                    0.009 (0.025)    -0.011 (0.015)       0.002 (0.002)
## knowl                          0.007 (0.025)   0.058*** (0.014)    -0.005** (0.002)
## trust_others                  0.048* (0.025)    0.032** (0.015)     -0.003 (0.002)
## Consistency                 -2.091*** (0.127) 1.494*** (0.073)
## Extremity                                                        0.035*** (0.002)
## satisfied_life                -0.047 (0.032)    -0.006 (0.018)      -0.002 (0.003)
## understand_issues             -0.013 (0.025)    -0.027* (0.014)      0.001 (0.002)
## too_complicated               -0.002 (0.025)     0.017 (0.015)      0.004** (0.002)
## tried_influence                0.023 (0.024)     0.019 (0.014)      -0.002 (0.002)
## Action_Score                  -0.016 (0.013)    -0.009 (0.007)      -0.001 (0.001)
## census_region                  0.010 (0.011)     0.003 (0.007)       0.001 (0.001)
## year                           0.001 (0.004)    0.006** (0.002)   -0.001*** (0.0004)
## str_PARTY                      0.031 (0.025)    -0.006 (0.014)       0.001 (0.002)
## PARTY_ID                       0.008 (0.006)    0.007** (0.003)     -0.001* (0.001)
## Constant                      -0.089 (8.387)   -12.204** (4.854)   3.154*** (0.738)
## --------------------------------------------------------------------------------
## Observations                       7,662             7,662             7,662
## R2                                 0.041             0.061             0.059
## Adjusted R2                        0.037             0.058             0.055
## Residual Std. Error (df = 7631)    0.984             0.570             0.087
## F Statistic (df = 30; 7631)      10.838***         16.592***         15.924***
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

These models are technically useable (significant F-statistic), they suffer from extremely low R^2 values. Though this model is barely useful, key trends and common knowledge are already visible, providing plenty of room for improvement by incorporating machine learning techniques.

Political knowledge is calculated as a single indicator variable if the respondent correctly answered which party was in the majority in the house, demonstrating a baseline understanding of current political conditions in the United States. About 60% of ANES respondents correctly answered this question, so it is a reasonably good cutoff question for “knowledge” within the purpose of this paper.

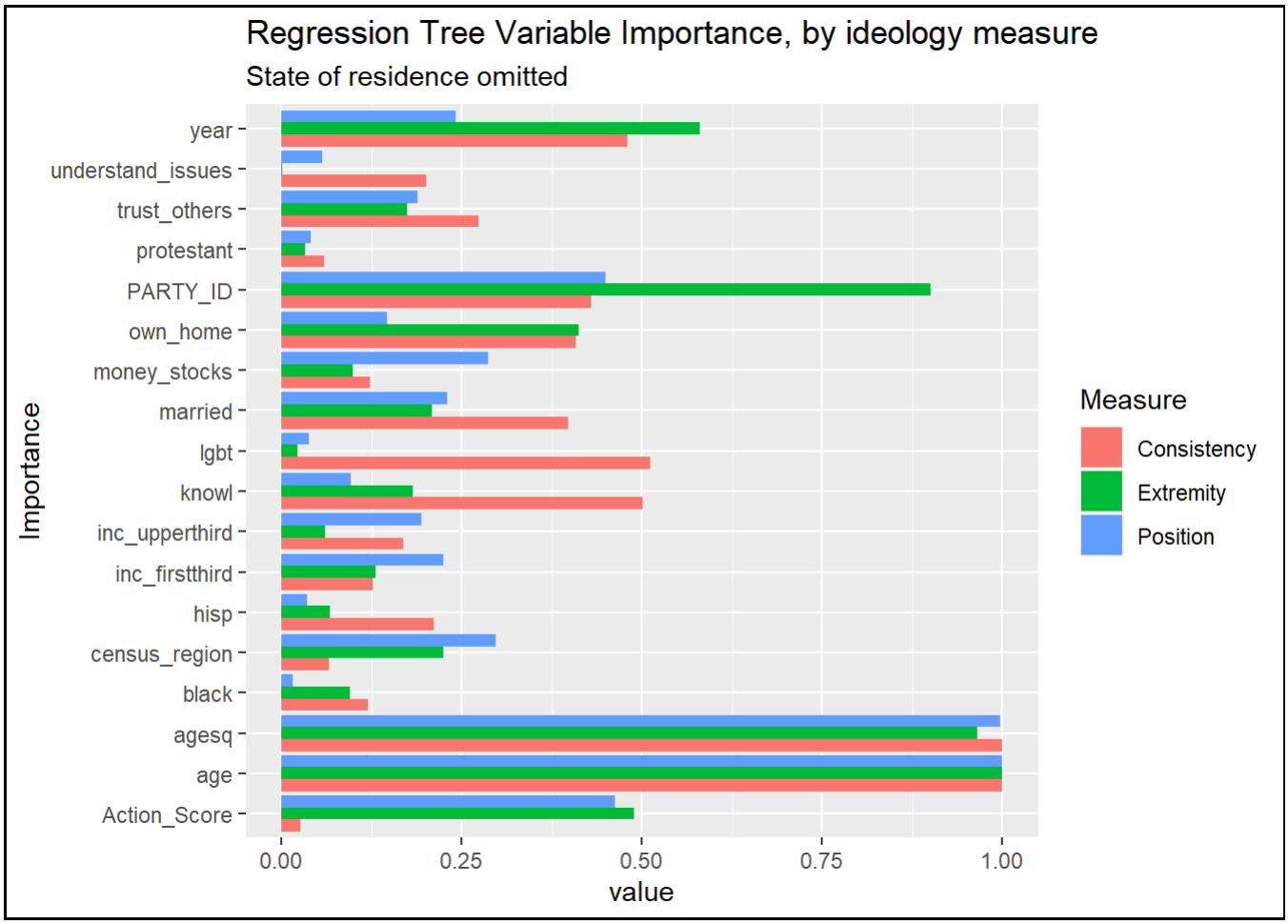
PARTY_ID is a question indicating a respondent’s partisan identification and strength of identification, with 4 being an independent. Relatively strong Republicans (PARTY_ID6) indicate lower ideological consistency but higher levels of extremity than their corresponding strong Democratic partisans (PARTY_ID2).

Regression Tree

The OLS regression above is useful in providing a sense of estimate magnitude and significance in a diagnostic sense, but abandoning the need for specific estimators may yield stronger predictive value from these variables. A regression tree does not provide nice coefficients, but inherently allows for complex interactions between all variables in order to achieve the best predictive performance.

It is immediately obvious that introducing a tree model increases the predictive capacity of this model dramatically. However, this should be taken with a grain of salt, since the premise of the model is to explode the number of terminal nodes, and thus predictive accuracy. I pruned these trees at the .0015 level. The traditional “rule of thumb” measure yielded 0 splits, meaning the “best” tree by minimizing error was essentially guessing randomly.

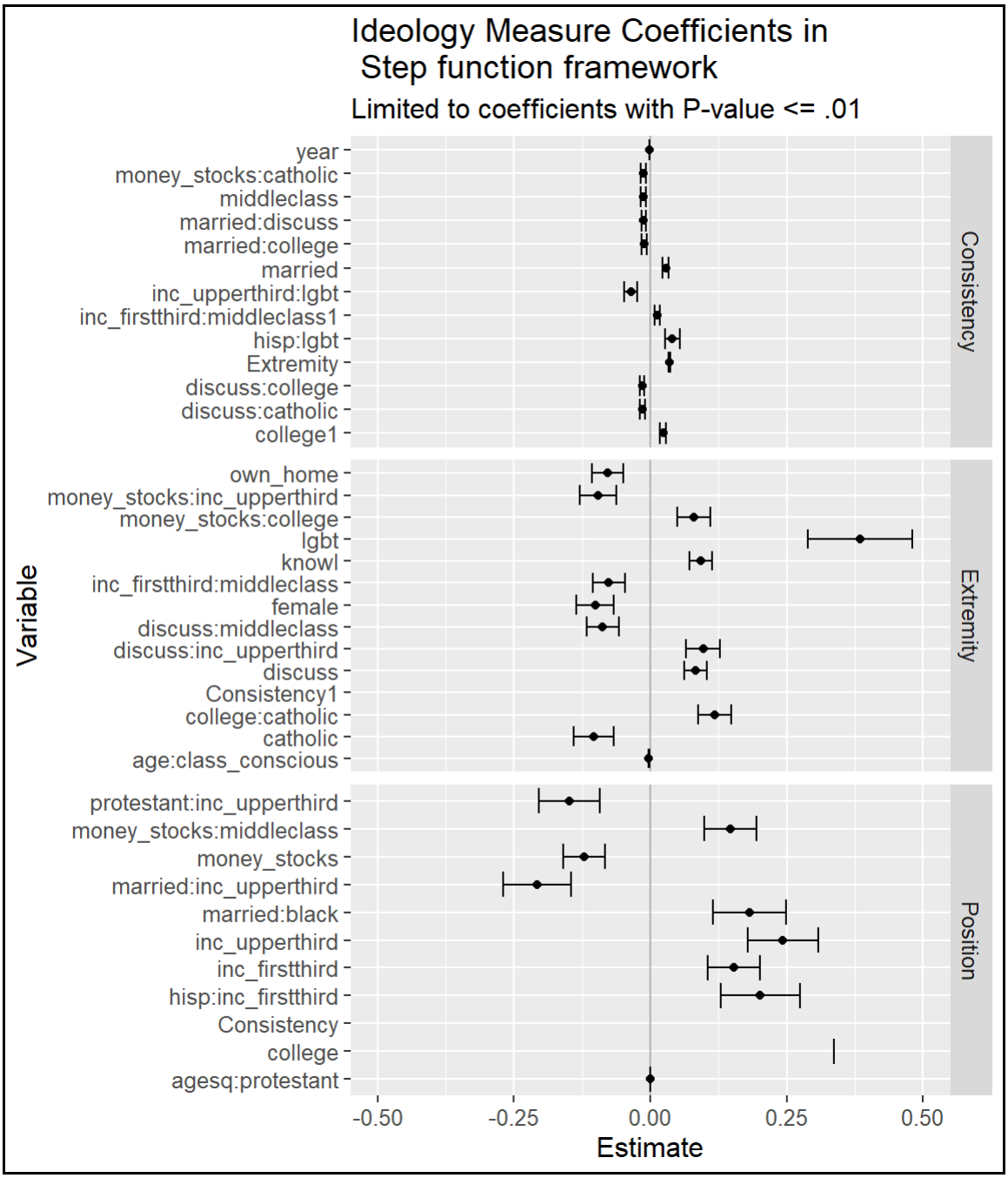
By imposing ~20 splits onto the tree, the R^2 of the tree improves dramatically, and we can pull variable importance plots from the model.



These plots do not communicate any directionality, rather the importance in generating “splits” in the tree. To put another way, when predicting the ideology of a given respondent, higher importance values indicate that splitting on that variable will yield a better prediction.

Step Regression

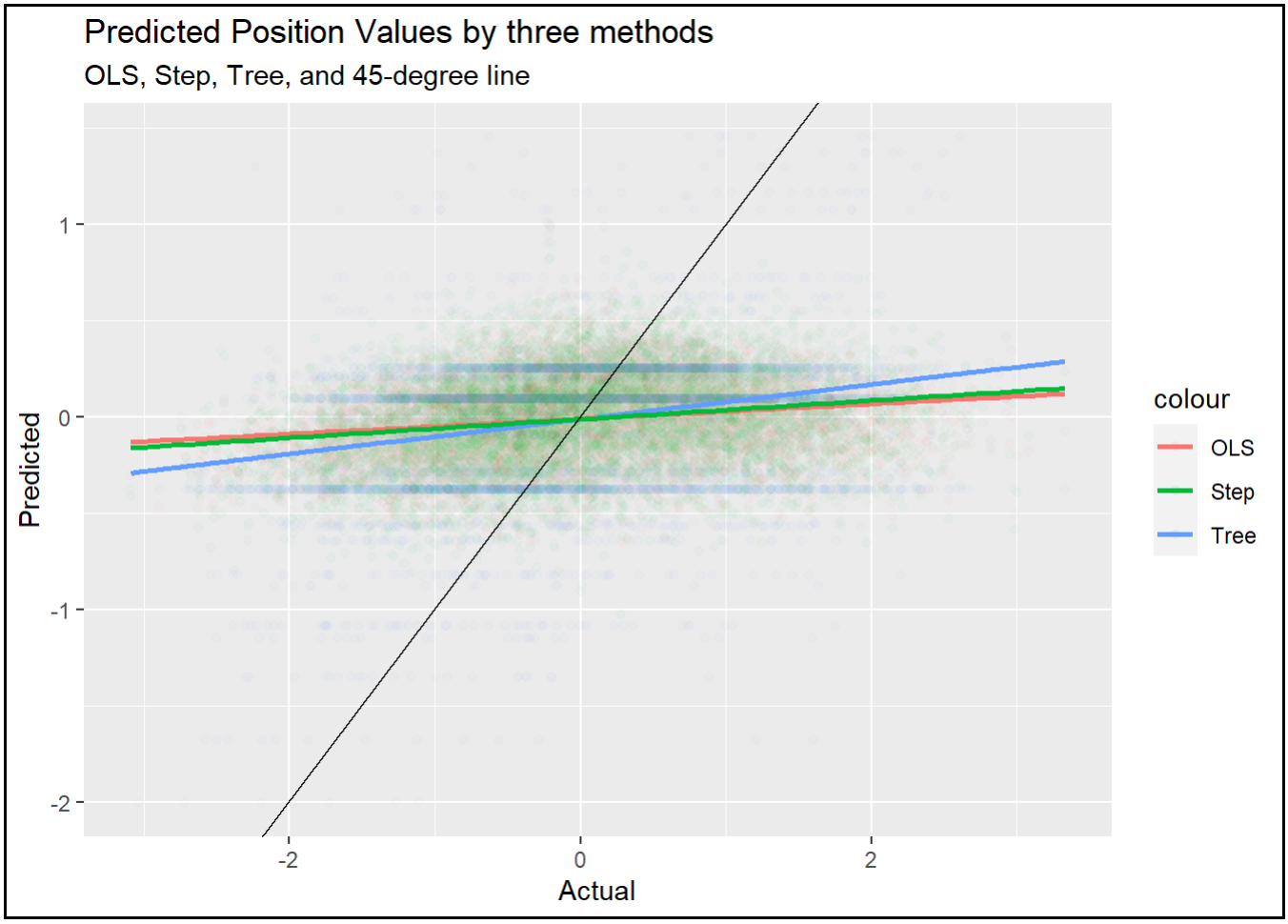
As a final model, splitting the difference between these two, is a stepped linear regression. This process runs dozens or hundreds of regressions, adding and subtracting variables & interactions to achieve the best fit. This process will not lead to a clean OLS model as seen in political science or economics literature, but can capture the "messy" identity-based interactions a simple additive model fails to capture.



The step regression splits the difference between the tree model and the OLS model in terms of predictability and is less of a "black box" for interpreting outcomes. The downside is reporting coefficients, since it generates potentially hundreds of interaction terms. For reporting, I omit all non-significant variables below the .01 level.

Introducing an algorithmically-selected set of interactions yields both a better model fit and new insight into ideological position. In some cases, the initial significance of a variable disappeared when interacted with one of the others, like hispanic identity in all three models.

Due to resource constraints, only one degree of interaction was available (i.e. all variables only interacted once). The step regression method could be applied to n-degrees with enough computing power and observations to make the model valid.



R-Squared Values for each regression

	Position	Extremity	Consistency
OLS	0.0408657	0.0612337	0.0589132
Tree	0.0676044	0.0912764	0.0640238
Step	0.0554071	0.0773703	0.0725597

Finally, comparing R^2 across each of these models, we see the step regression does strictly better than OLS, which is expected due to the nature of adding terms to a regression. Interestingly, it also outperforms the tree models in the case of the "consistency" variable. This implies the hierarchical structure of a tree model may actually be less explanatory than the additive linear structure of a regression.

The step model slightly outperforms OLS to a degree where the actual "fit" of the model may not be much better. However, from the step regression coefficients model, we see the process of algorithmically selecting interaciton terms isolated effects not immediately present in the first OLS model. Looking at the Extremity variable, the class and income position affected the predited Extremity of those who often discuss politics, for example.

Conclusion

This paper shows that ideology is deeply related not only to macro-identites like race, age, and religion, but interactions between those identities into more specific configurations has a real effect on ideological position, extremeness, and consistency. I also present a configuration of ideological structure which relies on current literautre but provides a basis for deeper analysis. By analyzing ideological consistency and position in this manner, direct estimates of position can be estimated against shocks.

Developing a highly-interacted methodology in demographic analysis of politics is critical as identities fractally expand under current political and social circumstances. This presents a major drawback to survey methods which rely on crosstabs as a basis of analysis of this depth. Application of this method to such surveys may uncover insights into novel opinion formations in the population.

Though not causal, a key use of the tree & step methods can provide a way to impute data into “Don’t Knows” in otherwise gold-standard surveys to provide a basis for building more robust datasets. In this analysis, a significant number of variables had to be dropped due to a single NA, which breaks regression and tree algorithms. For example, if one set of observations has NA’s in one variable, and another set has NA’s in another variable, *all* observations with either variable must be dropped in regression analysis. Or, both of those variables must be dropped. In a situation where NA’s can be assumed to be random, a methodology like this would provide a way to impute those observations to build out richer data.

Looking forward

A methodology like this requires a longer timeline of thought and participation, but I have demonstrated a basis for a deeper conversation for ideological structure and its roots in identity. Whether, or how, these interactions “cause” ideology is not clear, especially in the case of consistency

The most glaring issue with this setup is the structure of the ANES data. First it is not a panel dataset nor is it balanced, so these estimates may not be valid on that basis alone. There is no causal claim that can be made from this per se, but introducing a linked-panel element in a broader study could yield interesting results in the field of studying ideology and ideology formation. Similarly, the presence of Refuse/Don’t Know’s complicates any survey study, since they cannot necessarily be considered randomly distributed and computed away.

Future avenues of research include creating a node-edge graph of a wider array of ideological components, tracking centrality among partisans and non-partisans, extremists and non-extremists. Measures of “closeness” can be translated into an edge list and therefore a graph of ideology with key, or central, beliefs in the center.