301-2 Final Project

John Lee

March 14, 2019

Introduction

My proposed data source is the General Social Survey (GSS), which is run by the National Opinion Research Center (NORC) at the University of Chicago. The GSS is one of the best known sources of nationally representative public opinion survey data in the U.S., with data that dates back to the 1970s. The GSS uses a stratified probability-based sampling method and asks a large number of questions related to social, cultural, political, and economic issues. The survey is typically done once every two years.

Here is a link to the dataset: https://gssdataexplorer.norc.org/ The website has a user-friendly platform from which we can just pick and choose which variables we would like to download ("GSS Data Explorer"). My main reason for choosing this dataset is that for many relevant topics, the survey has been asking the same questions for the past 40+ years.

My dependent variable of interest is the subject's answer to a question about party ID. It asks the following: *"Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?" The answer choices are organized using a logical and intuitive 7-point scale from "Strong Democrat" to "Strong Republican."*

The citation for the data source is the following:

Smith, Tom W, Michael Davern, Jeremy Freese, and Michael Hout. General Social Surveys, 1972-2016 [machine-readable data file] /Principal Investigator, Tom W. Smith; Co-Principal Investigator, Michael Davern; Co-Principal Investigator, Jeremy Freese; Co-Principal Investigator, Michael Hout; Sponsored by National Science Foundation. –NORC ed.– Chicago: NORC at the University of Chiago [producer]; Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [distributor], 2018.

I have two research questions:

(1) What explains support for political parties in the U.S. context? In particular, I'll focus on support for the Democratic Party (because out of the two major parties, the letter D comes before R). The focus here is on identifying statistically and substantively significant predictors (i.e., which specific predictors matter? And how much do these individual predictors matter?)

(2) Overall (i.e., multiple relevant predictors are used), how well can we predict support for political parties? How good is the "best" model? The focus here is on assessing the adjusted R-squared and the size of the test MSE.

My research question is necessarily predictive, because I don't have the kind of data I need (e.g., panel data) for a more robust method that provide a good basis for causal inference (e.g., linear FE models with robust standard errors). However, because most people agree that policy outcomes are largely shaped to public preferences – and that preferences are in some way tied to party ID and partisan messaging – I think a less ambitious project based on predictive goals would still have some merit.

Overview of the project report:

(1) Exploratory Data Analysis (EDA)

(2) Model-building

(3) Model Validation

(4) Model Testing

Part 1: EDA

During the data collection stage, my goal was to find all of the variables that were available between 1974-2016. The full sample had 62,466 observations and 64 variables (including DV and year).

The data cleaning process entailed doing the following (please refer to my data cleaning R script for more details):

- Filter in variables of interest (remove redundant predictors, predictors with a lot of missing data)
- Rename predictor names (more intuitive names), set the correct variable type (e.g., convert character variables into factor variables)
- Recode the values (e.g., sometimes needed to reverse code); this entailed creating custom functions
- For the categorical variables, explicitly set the reference categories (e.g., New England for geographic region)

The final dataset had 17,718 observations and 35 variables (including the DV and year). The DV is called `strong_dem`. It is measured using a 1-7 scale, where 1 indicates that the respondent self-identifies as a strong Republican and 7 indicates that the respondent identifies as a strong Democrat. I decided to operationalize the DV as a continuous (or numeric) variable because the distance between each answer choice is roughly the same, the answer choices are symmetrical (e.g., strong Democrat to Strong Republican), and having a continuous DV would let me use linear modeling techniques – which are generally easier to interpret.

Below is a list of the variables in the dataset. A clarifying note about the variables: there are many variables that look something like this `moresp_natarms`. These variables measure the respondent's policy preferences regarding a set of specific issues: e.g., welfare, public education, foreign aid, and so on. These predictors are binary variables: a 1 indicates that the respondent wants more public spending in this area (e.g., on public education); a 0 indicates that the respondent does not favor more spending (e.g., keeping spending the same or reducing it).

```
##  [1] "strong_dem"           "year"                 "age"
##  [4] "black"                "race_text"            "female"
##  [7] "num_kids"             "num_sibs"             "hh_size"
## [10] "religion"             "weekly_relig_attend"  "fundamentalist"
## [13] "ever_divorced"        "married"              "subj_classid_text"
## [16] "ln_famincome"         "ft_job"               "yrs_educ"
## [19] "dad_yrs_educ"         "mom_yrs_educ"         "lwparents_at16"
## [22] "geomobile_at16"       "rural_restype_at16"   "geo_region"
## [25] "moresp_natforeignaid" "moresp_natarms"       "moresp_natcrime"
## [28] "moresp_natdrug"       "moresp_nateduc"       "moresp_natenvir"
## [31] "moresp_natwelfare"    "moresp_nathealth"     "moresp_natrace"
## [34] "moresp_natspace"      "pol_liberalism"
```

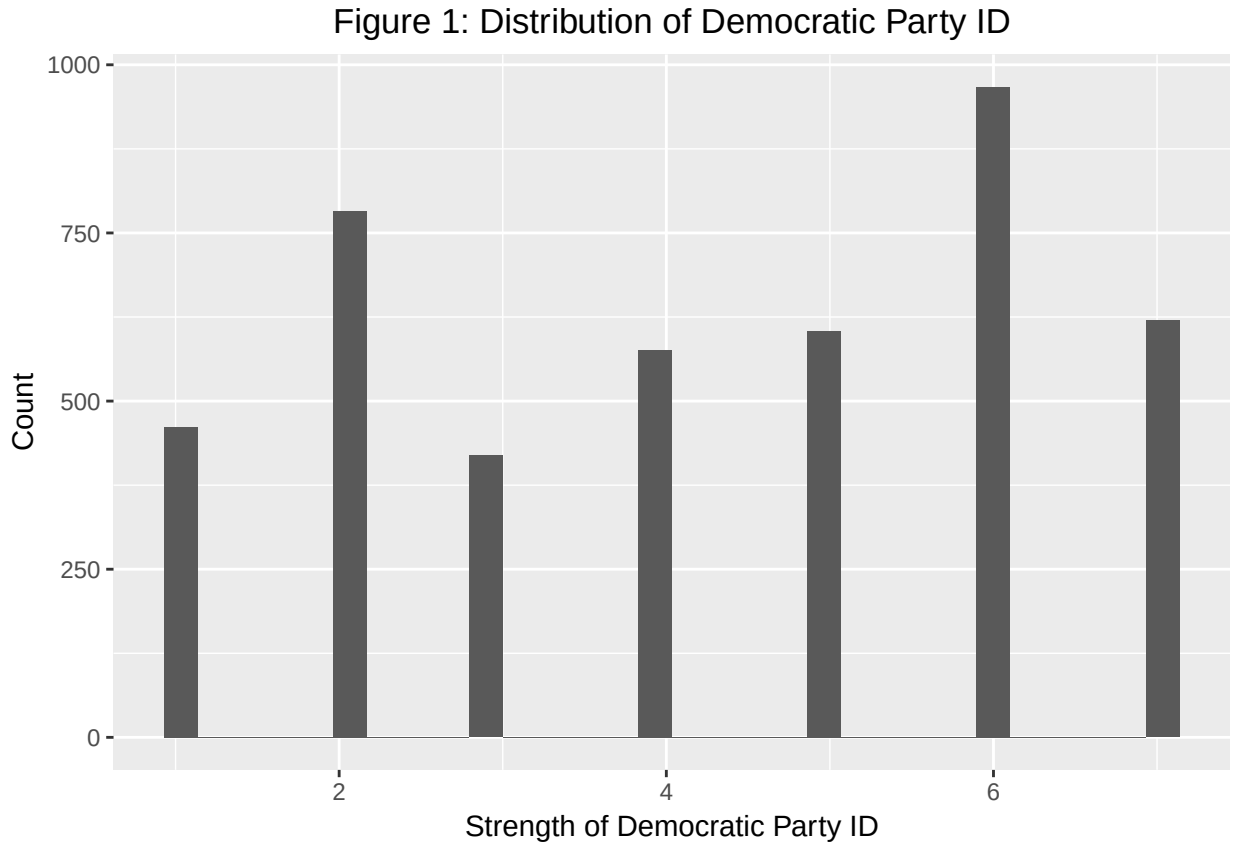
Before doing any of the analysis, I randomly split the final dataset into four parts (with the seed set to 3, per class norms):

(1) EDA: 25%
(2) Model-building: 25%
(3) Model validation: 25%
(4) Model testing: 25%

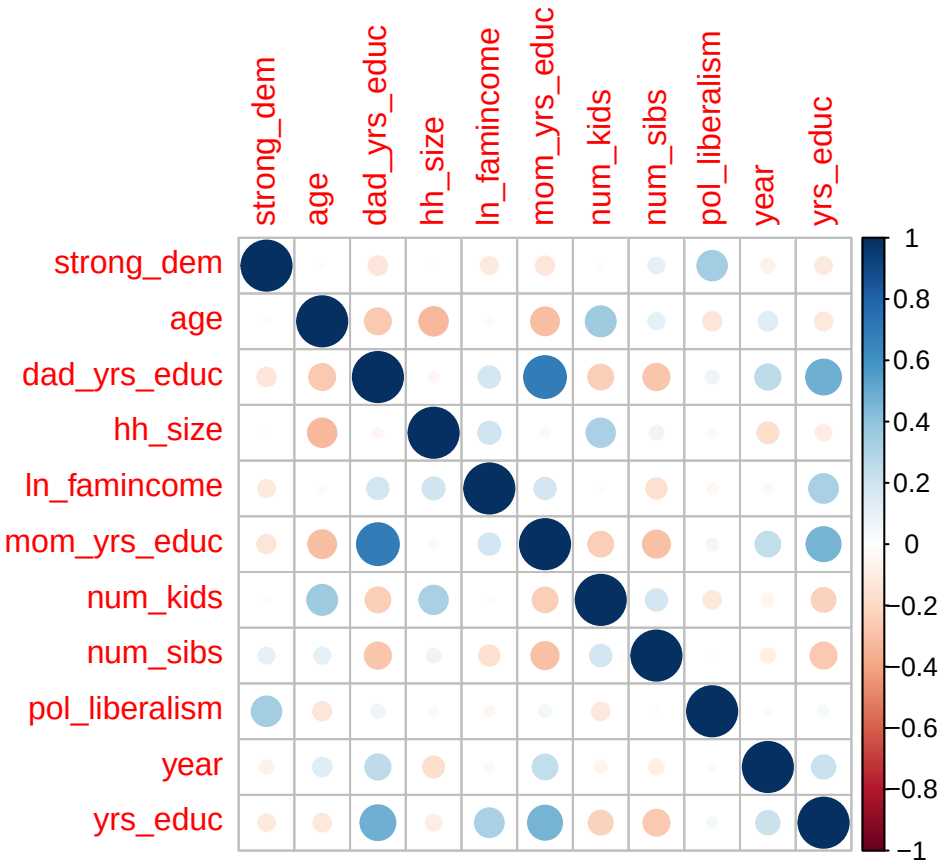
With the EDA dataset, I'll perform a basic EDA: e.g., look at the distribution of the DV, examine the bivariate relationships. With the model-building dataset, I'll select the candidate models: i.e., the best model for each method (e.g., lasso, ridge, PCR). With the model validation set, I'll compare the candidate models and choose the "best" model as my finalist. Finally, I'll evaluate the performance of my final model using the held-out test set.

Now, I'll proceed to the EDA. Figure 1 below displays the distribution of the DV: i.e., a 1-7 point scale indicating how strongly the respondent supports the Democratic Party. There is a large number of respondents

at each level of the DV, which is good for estimation purposes.

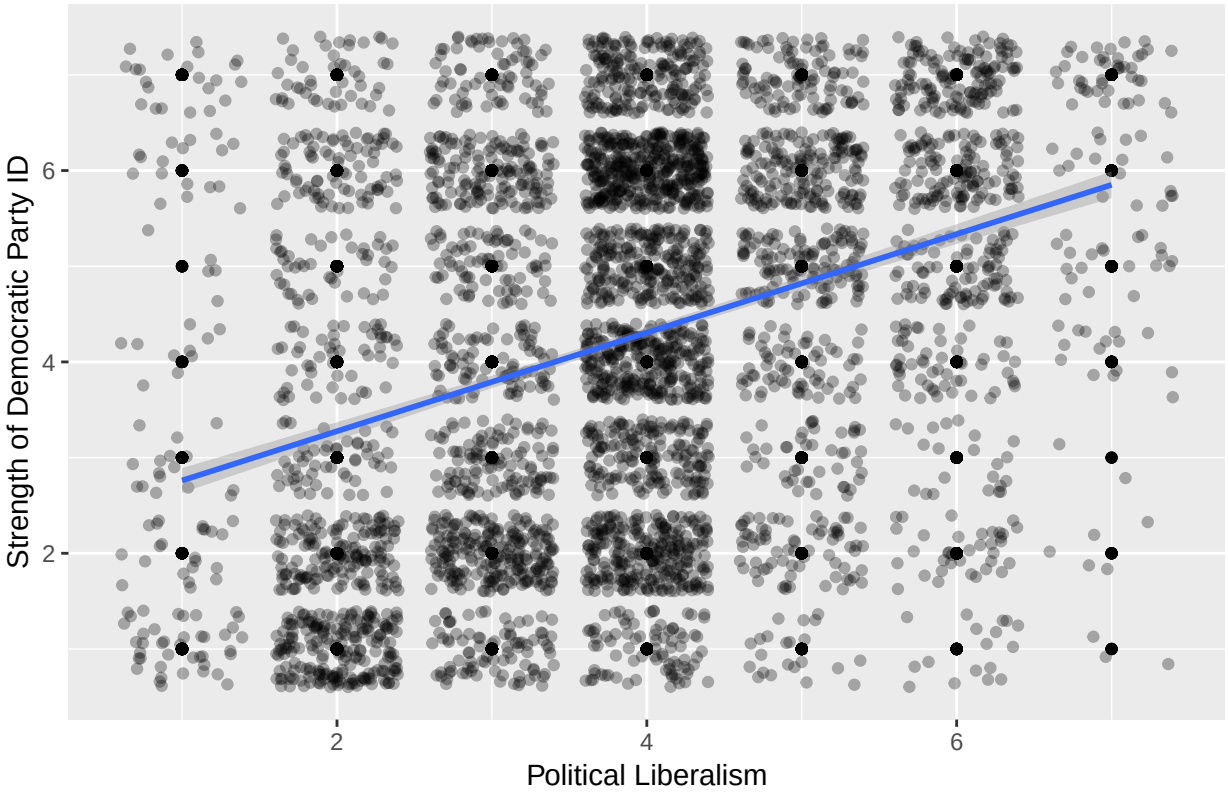


Below is a correlation matrix which visually depicts the strength of the linear association among the numeric variables. We can see most of the other variables are only weakly correlated with the DV. The one exception is political liberalism, which has a moderately positive association with the strength of support for the Democratic party.



To investigate this further, I create a scatter plot of the DV against political liberalism. Indeed, it looks as though there is a positive correlation between the two variables.

Figure 2: Scatter Plot of Democratic Party ID and Pol. Liberalism



Next, I also examine the relationship between several categorical predictors and the DV. First, I'll look at two key demographic predictors: gender and race. According to the results below, it looks as though women (relative to non-women) and non-whites (relative to whites) are stronger supporters of the Democratic Party.

Figure 3: Democratic Party ID and Gen

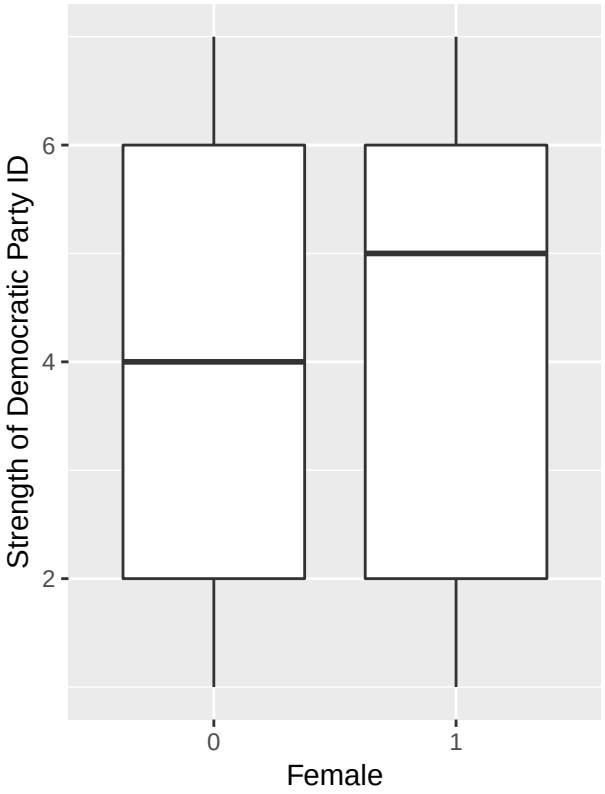
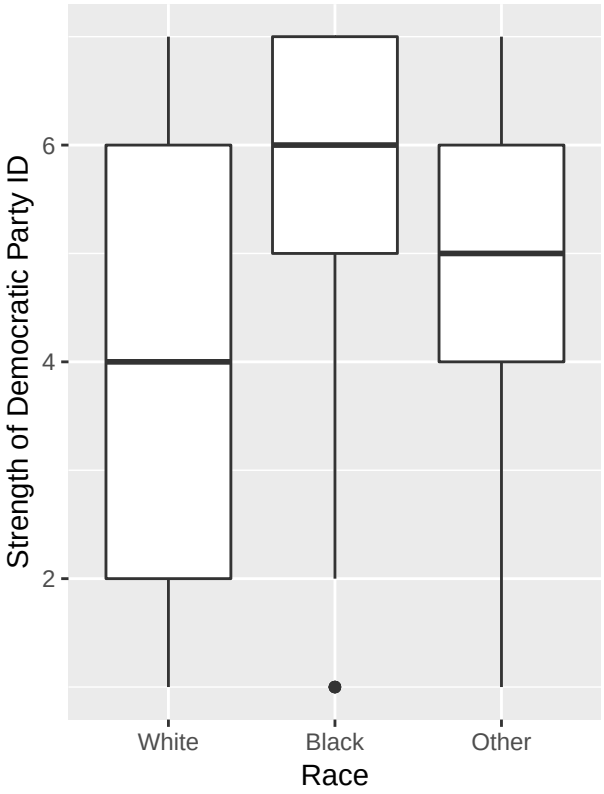
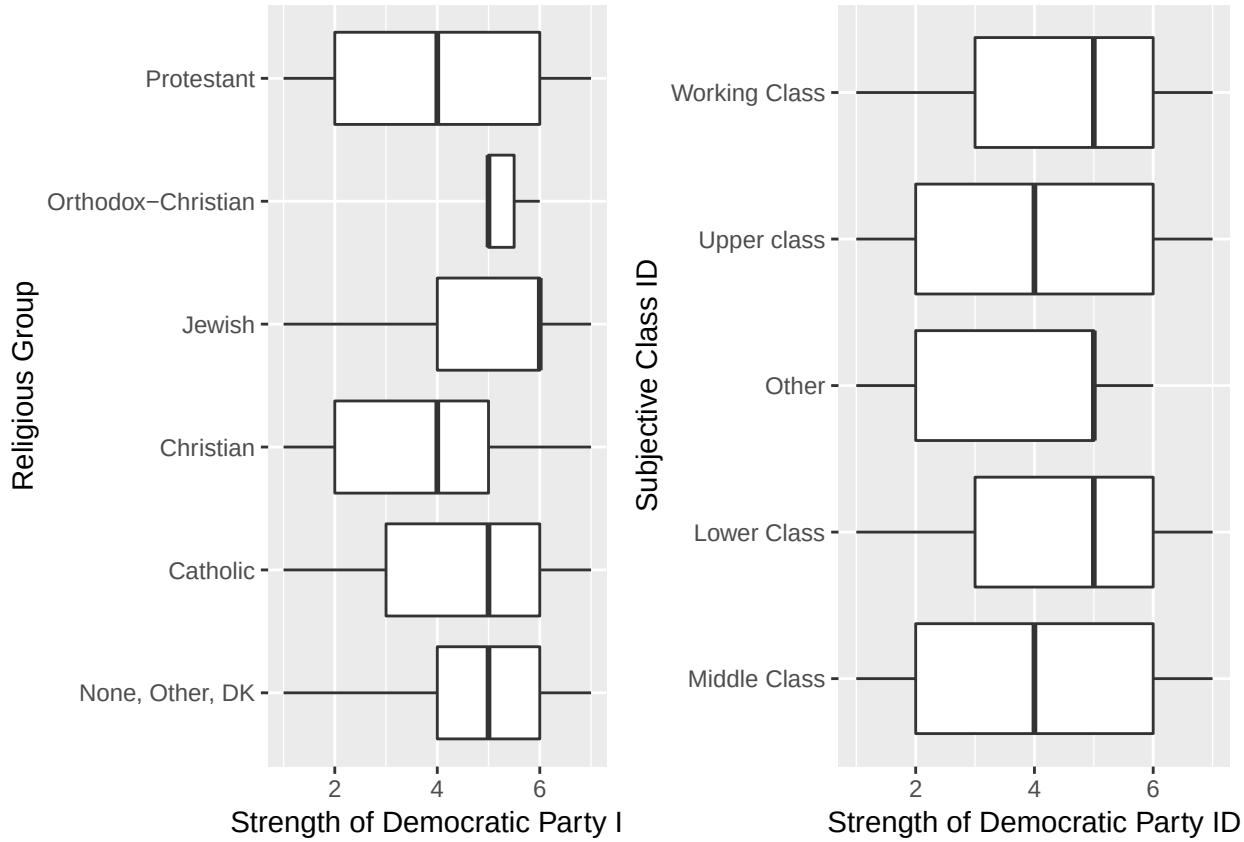


Figure 4: Democratic Party ID and Rac



Next, I'll look at the relationship between the DV and two social factors: subjective class ID and religion. The results below suggest that the divisions by class and religion are meaningful. Those who identify as members of the lower and working classes tend to be more aligned with the Democrats; in contrast, those in the middle and upper classes are evenly distributed between both parties. Protestants tend to be split among the parties, but those in other religious groups lean more toward the Democrats.



While the results above are certainly not exhaustive, and I could have examined more sets of bivariate relationships, the points are clear: many of the predictors – especially the categorical variables – seem to be correlated with the DV. However, what's less clear from the EDA alone is how well we can predict the DV using these predictors. Which predictors are the most important? To answer this question, I proceed to the model-building stage of my report.

Part 2: Model-building

In the model-building stage, my goal is the following: I want to identify the best model associated with each method (e.g., OLS, ridge regression, lasso regression, PCR). These best models will together constitute my candidate models – and they'll be advanced to stage 3 (validation), where they'll be compared against each other and a finalist will be chosen.

The model-building stage as two parts. In the first part of this stage, I'll develop five OLS models based on my knowledge of the existing literature as well as the EDA in the previous section. I'll fit each of these models using one half of the model-builing set; then, I'll test them using the second half of the model-building set. The model with the smallest test MSE will become one of my candidate models.

In the second part of the model-building stage, I'll use k-fold cross-validation (with k = 10) to identify the optimal tuning parameters for each of the methods used: i.e., lambda for ridge and lasso regression; and M (or the number of principal components) for PCR and PLS. The model built using the optimal tuning parameter will be the best model from each method – and these best models will become the rest of the candidate models.

Best OLS Model

Below, is the baseline OLS model. Each of the five OLS models I will test adds a single interaction term to the baseline model (also displayed below). The purpose of each interaction term is to test whether the link between the predictor (e.g., race) and support for the Democrats varies as a function of year. There are reasons to expect this would be the case. For example, many pundits and scholars have argued that the U.S. public has become more polarized in recent years – i.e., the political parties have allegedly become more ideologically pure and homogeneous. If this were the case, then we would expect the interaction term for `year:pol_liberalism` to be statistically significant and positive.

```
# Create the baseline fmla
baseline_fmla <- "strong_dem ~ year + age + race_text + female + religion + subj_classid_text +
ln_famincome + yrs_educ + pol_liberalism + moresp_natwelfare + moresp_natdrug + moresp_natarms +
  moresp_natenvir + moresp_natrace "

# Mod names
mod_names <- c("+ year:race_text",
               "+ year:female",
               "+ year:ln_famincome",
               "+ year:yrs_educ",
               "+ year:pol_liberalism")
```

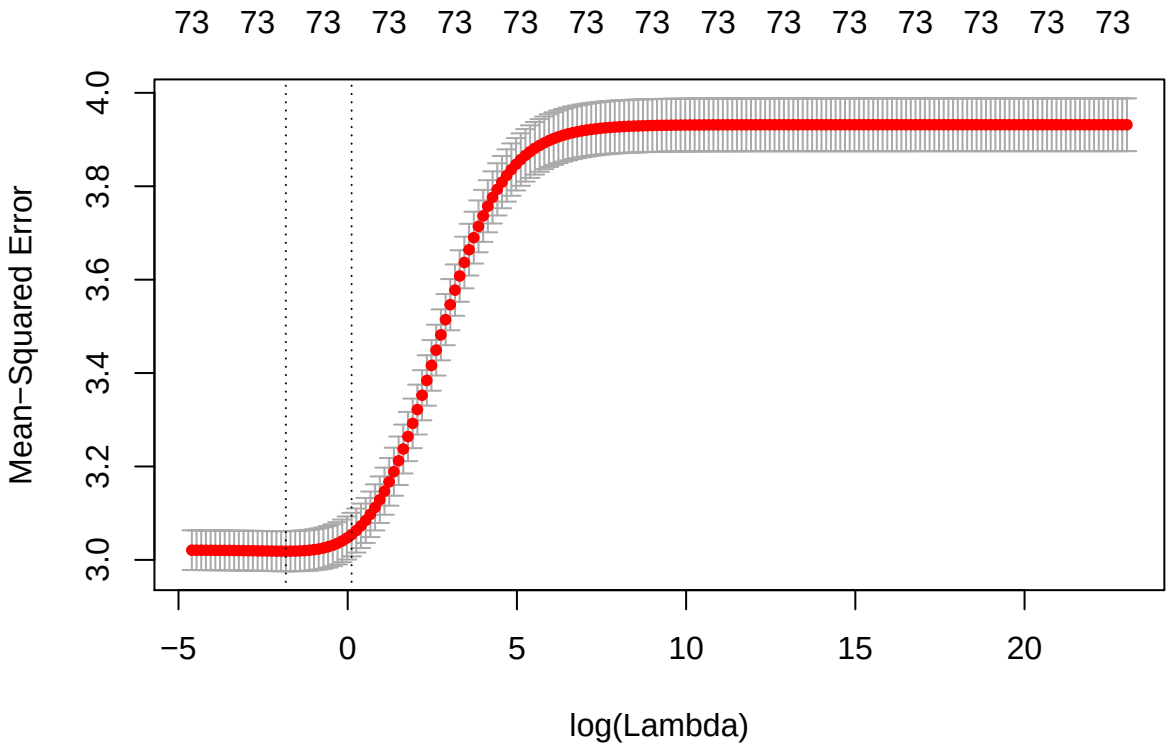
I've displayed the findings of the OLS analysis below. The results indicate that the model that includes the interaction term `year:pol_liberalism` generated the lowest test MSE. For that reason, this will become my best OLS model and advance as one of the candidate models. However, it's worth noting that the differences in test MSE are quite small, and that given a slightly different dataset (e.g., due to a different seed), another OLS model could have achieved the lowest test MSE.

mod_names	test_mse
+ year:pol_liberalism	2.957
+ year:yrs_educ	2.960
+ year:race_text	2.968
+ year:female	2.972
+ year:ln_famincome	2.973

Next, I use 10-fold CV to identify the optimal tuning parameters for the ridge regression, lasso regression, PCR, and PLS, respectively.

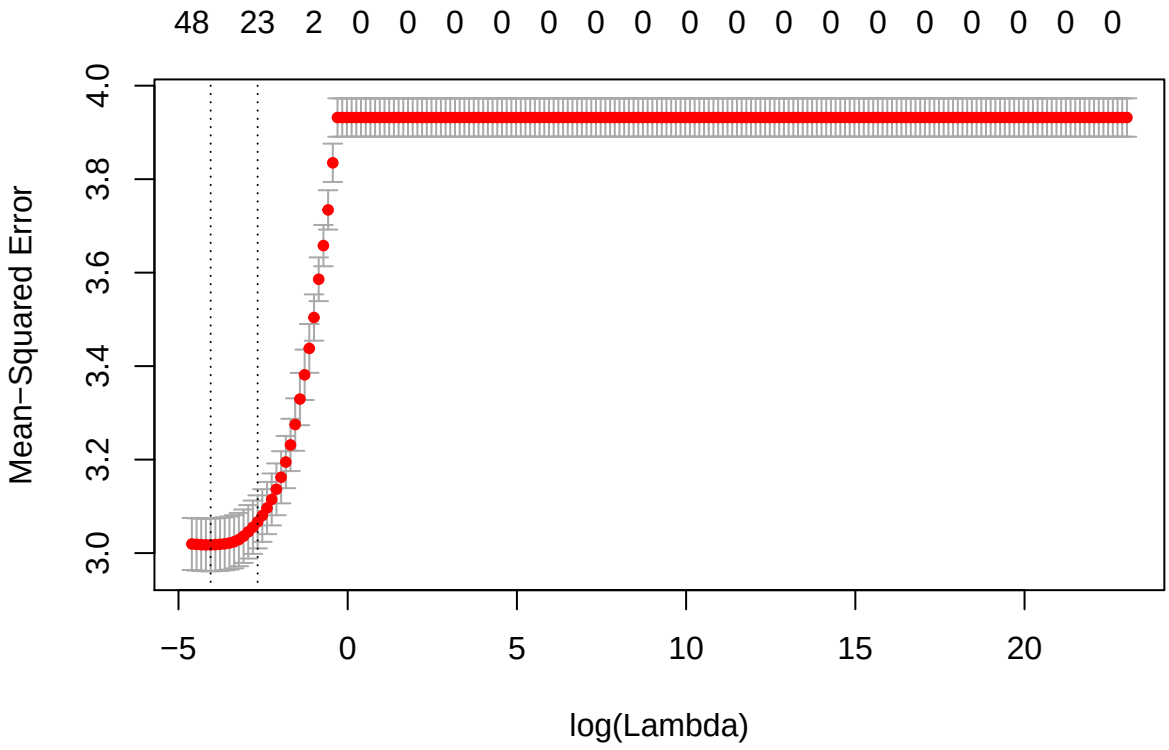
Best Ridge Regression Model

The plot below shows the MSE as a function of the logged lambda. According to the CV process, the ridge regression generates the lowest MSE when the lambda is 0.1607053. The largest lambda value within one SD of the minimum MSE is 1.122668.



Best Lasso Regression Model

Next, I move on to the lasso regression model. Unlike the ridge, the lasso actually zeros out the coefficients of predictors that are insufficiently associated with the DV. The plot below shows the MSE as a function of the logged lambda. According to the CV process, the lasso regression generates the lowest MSE when the lambda is 0.01742633. The largest lambda value within one SD of the minimum MSE is 0.0698588.



Below, I fit the ridge and lesso regression models using the optimal tuning parameters identified in the CV process above: i.e., the lambda values that generated the lowest test MSE as well as the largest lambda values with test MSEs within one SD of the minimum MSE. The regression coefficients with each of these models are displayed below for comparison purposes.

In general, the slope coefficients seem quite similar across the models. What's especially interesting is that the lasso models seem to be zeroing out many of the coefficients – which indicates that many of the predictors are not meaningfully associated with the DV, once we've accounted for the other predictors. For example, it turns out that when other key predictors are included in the model (e.g., key demographic and social predictors), geographic region does not seem to add any additional predictive ability to the model.

name	ridge_min	ridge_1se	lasso_min	lasso_1se
(Intercept)	3.945	4.475	4.993	4.729
year	-0.008	-0.005	-0.006	0.000
age	0.007	0.003	0.005	NA
black0	-0.273	-0.263	-0.914	-0.917
black1	0.294	0.262	NA	NA
race_textWhite	-0.222	-0.194	-0.153	-0.016
race_textBlack	0.288	0.264	NA	0.001
race_textOther	-0.016	0.003	NA	NA
female0	-0.070	-0.058	-0.096	NA
female1	0.070	0.058	0.000	NA
num_kids	-0.009	0.001	NA	NA
num_sibs	0.022	0.020	0.021	0.014
hh_size	0.007	-0.003	NA	NA
religionNone, Other, DK	0.112	0.104	NA	NA

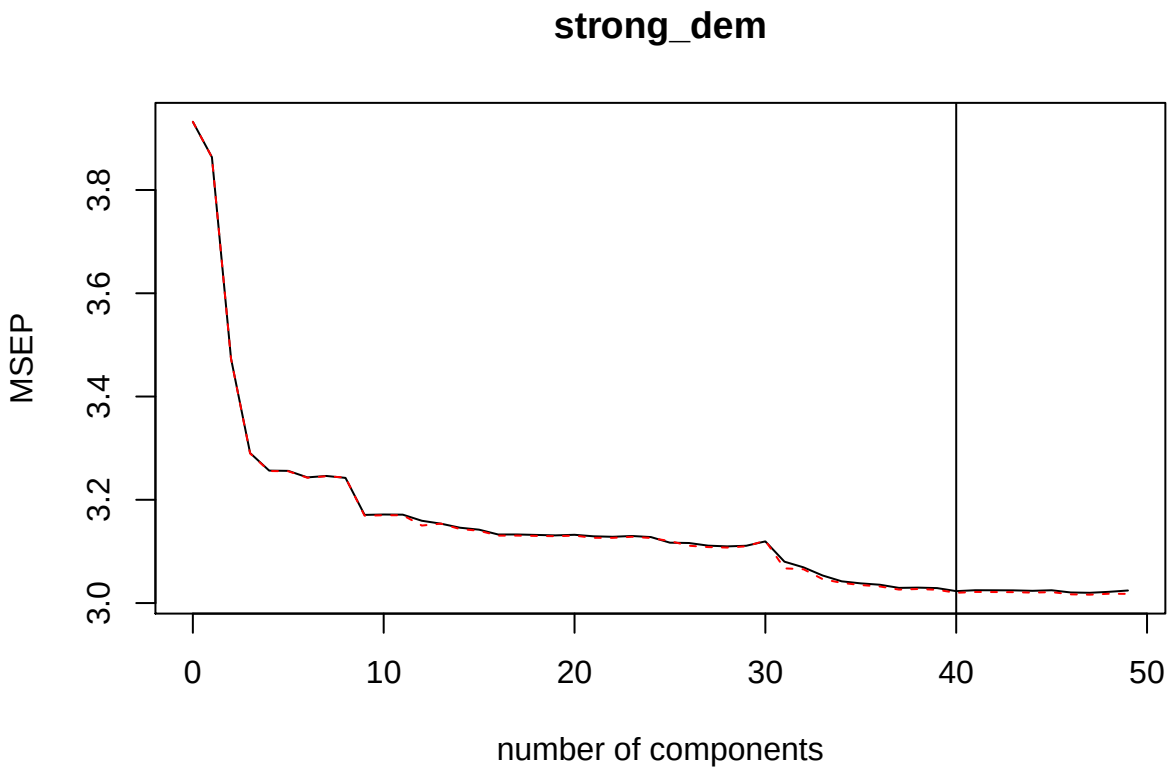
name	ridge_min	ridge_1se	lasso_min	lasso_1se
religionCatholic	0.195	0.135	0.063	NA
religionChristian	-0.107	-0.098	-0.057	NA
religionJewish	0.700	0.501	0.522	0.185
religionOrthodox-Christian	-0.055	-0.046	NA	NA
religionProtestant	-0.266	-0.196	-0.329	-0.216
weekly_relig_attend0	0.067	0.068	0.087	NA
weekly_relig_attend1	-0.067	-0.068	0.000	NA
fundamentalist0	-0.007	0.018	NA	NA
fundamentalist1	0.007	-0.018	NA	NA
ever_divorced0	-0.106	-0.081	-0.165	NA
ever_divorced1	0.106	0.081	0.000	NA
married0	0.029	0.030	0.023	NA
married1	-0.029	-0.030	0.000	NA
subj_classid_textMiddle Class	-0.042	-0.038	NA	NA
subj_classid_textLower Class	-0.069	-0.044	NA	NA
subj_classid_textOther	-0.197	-0.091	NA	NA
subj_classid_textUpper class	-0.568	-0.401	-0.479	-0.227
subj_classid_textWorking Class	0.131	0.100	0.158	0.069
ln_famincome	-0.033	-0.034	-0.026	-0.011
ft_job0	-0.028	-0.013	-0.001	NA
ft_job1	0.028	0.013	NA	NA
yrs_educ	-0.004	-0.008	-0.001	NA
dad_yrs_educ	-0.022	-0.019	-0.023	-0.023
mom_yrs_educ	-0.024	-0.022	-0.025	-0.026
lwparents_at160	0.017	0.012	NA	NA
lwparents_at161	-0.017	-0.012	NA	NA
geomobile_at160	0.079	0.057	0.122	NA
geomobile_at161	-0.078	-0.057	0.000	NA
rural_restype_at160	-0.006	0.002	NA	NA
rural_restype_at161	0.006	-0.002	NA	NA
geo_regionNew England	0.013	0.048	NA	NA
geo_regionMiddle Atlantic	-0.114	-0.047	-0.074	NA
geo_regionMountain	-0.059	-0.062	-0.024	NA
geo_regionNE Central	-0.067	-0.051	-0.048	NA
geo_regionNW Central	0.091	0.062	0.024	NA
geo_regionPacific	-0.002	0.004	NA	NA
geo_regionSE Central	0.140	0.084	0.068	NA
geo_regionSouth Atlantic	0.044	0.013	NA	NA
geo_regionSW Central	0.074	0.040	0.017	NA
moresp_natforeignaid0	-0.048	-0.045	-0.019	NA
moresp_natforeignaid1	0.047	0.045	0.000	NA
moresp_natarms0	0.090	0.090	0.138	0.029
moresp_natarms1	-0.089	-0.090	NA	0.000
moresp_natcrime0	0.012	0.006	NA	NA
moresp_natcrime1	-0.012	-0.006	NA	NA
moresp_natdrug0	-0.088	-0.082	-0.159	-0.113
moresp_natdrug1	0.088	0.082	0.000	0.000
moresp_nateduc0	-0.093	-0.084	-0.158	-0.063
moresp_nateduc1	0.093	0.084	0.000	0.000
moresp_natenvir0	-0.066	-0.074	-0.096	-0.026
moresp_natenvir1	0.065	0.074	0.000	0.000
moresp_natwelfare0	-0.154	-0.154	-0.286	-0.233

name	ridge_min	ridge_1se	lasso_min	lasso_1se
moresp_natwelfare1	0.153	0.154	NA	NA
moresp_nathealth0	-0.151	-0.145	-0.291	-0.265
moresp_nathealth1	0.150	0.145	NA	NA
moresp_natrace0	-0.143	-0.147	-0.271	-0.250
moresp_natrace1	0.142	0.147	NA	NA
moresp_natspace0	0.025	0.030	0.016	NA
moresp_natspace1	-0.025	-0.030	NA	NA
pol_liberalism	0.373	0.257	0.404	0.400

Best Principal Component Regression Model

Next, I'll find the optimal value of M (or the number of principal components) for a first-order PCR model. PCR (and PLS) can be useful tools when many of the predictors are highly correlated. PCR can be used to create more parsimonious models by converting the original predictors into a set of principal components which efficiently represent the variance of the original predictors.

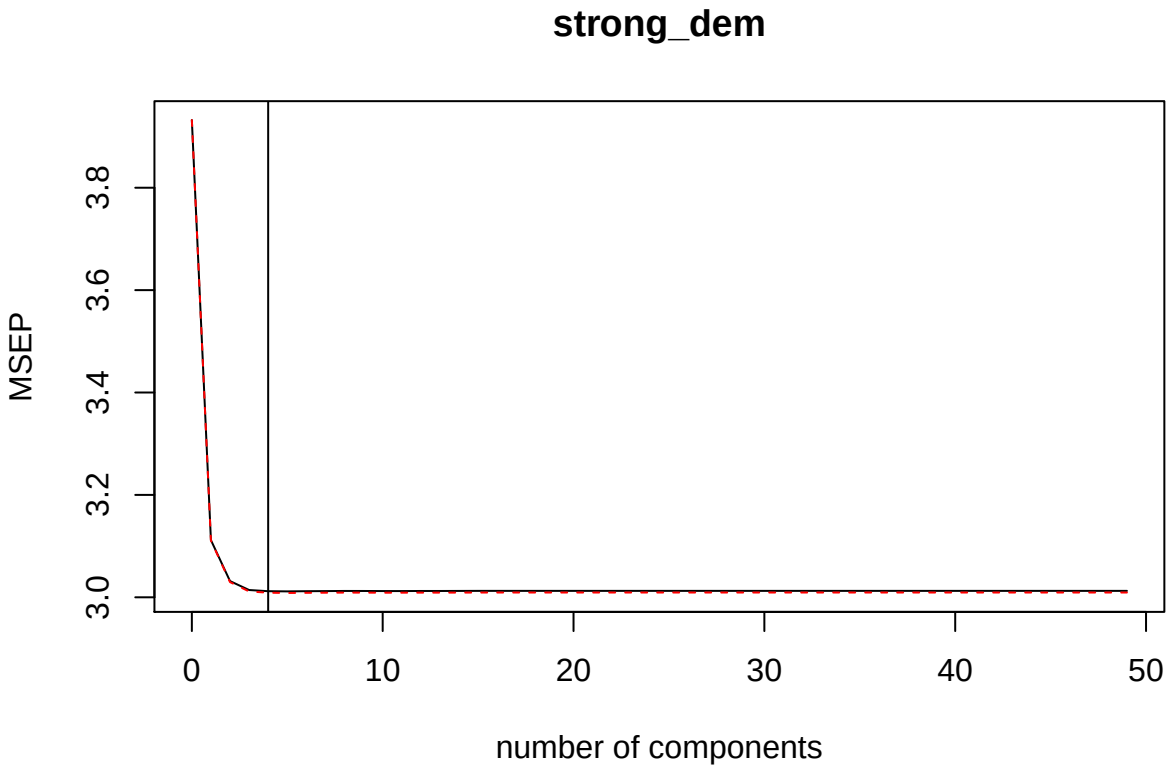
As illustrated via the plot below, for PCR, further reductions in MSEP appear to decline after M = 40 or so. As such, the PCR model with M = 40 is chosen as the candidate model for this method.



Best Partial Least Squares (PLS) Regression Model

Similarly, we can also use 10-fold CV to find the optimal M for a PLS regression model. As indicated by the plot below, the best M is around a 4. After 4 principal components, adding additional PCs does not appear to meaningfully improve the accuracy of the PLS model.

It's worth noting that the difference between the ideal M for the PCR and PLS models are quite large: i.e., 40 and 4, respectively. This is likely rooted in the fact that the PLS is a supervised form of PCR, in which only PCs that are related to both the original predictors and also the DV are selected. By imposing this additional constraint, it is often the case that the M of a PLS is smaller than that of a PCR. In this case, the optimal M of the PLS is substantially smaller (i.e., 4 compared to 40), which suggests that only a small number of principal components are related to the DV.



Best Forward Selection Model

Finally, I'll employ two additional methods: the forward and backward selection methods. Best subset selection was not used because it was too computationally demanding: it requires fitting and comparing 2^p models. I tried running it and my R crashed.

To select the optimal number of predictors, I'll again use 10-fold CV. When I use this CV process with forward selection, it appears as though the optimal number of predictors is 48. However, the CV MSE is very similar across the top 5 options: it appears as though 44-48 predictors could have been chosen. Since everything else being equal (or sufficiently similar), parsimonious models are preferred, I'll use the model with 44 predictors as my candidate model for the forward selection method.

```
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
```

```
## Reordering variables and trying again:
## Reordering variables and trying again:
```

model_index	test_mse
48	3.020
46	3.022
47	3.023
45	3.025
44	3.033

Best Backward Selection Model

Similarly, I repeat running 10-fold CV to identify the optimal number of predictors for backward selection. Again, it appears as though this occurs when 48 predictors are used. However, the differences in CV MSE for models that use 44-48 predictors are very similar. Thus, using the same rationale as the one described above, the backward selection model with 44 predictors will be used as one of my candidate models.

```
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
## Reordering variables and trying again:
```

model_index	test_mse
48	3.025
47	3.025
46	3.048
45	3.051
44	3.057

```
## Reordering variables and trying again:
## Reordering variables and trying again:
```

Since both the forward and backward selection candidate models use 44 predictors, are they the same 44 predictors? We can check this below. It turns out that while the models share most of the same predictors, there are a few differences – and consequently, the regression coefficients for the shared predictors are also slightly different.

name	fwd	back
(Intercept)	2.252	2.481
year	-0.007	-0.007
age	0.008	0.008
black1	1.080	1.087
female1	0.125	0.125
num_kids	-0.016	-0.017
num_sibs	0.026	0.026
hh_size	0.020	0.019

name	fwd	back
religionCatholic	0.402	0.391
religionChristian	0.048	0.052
religionJewish	0.904	0.906
religionOrthodox-Christian	0.090	0.097
weekly_relig_attend1	-0.156	-0.150
fundamentalist1	-0.053	-0.048
ever_divorced1	0.210	0.214
married1	-0.063	-0.062
subj_classid_textLower Class	0.000	-0.003
subj_classid_textUpper class	-0.588	-0.588
subj_classid_textWorking Class	0.198	0.197
ln_famincome	-0.041	-0.043
ft_job1	0.066	0.065
yrs_educ	-0.005	-0.005
mom_yrs_educ	-0.040	-0.041
lwparents_at161	-0.030	-0.030
geomobile_at161	-0.159	-0.162
rural_restype_at161	0.018	0.017
geo_regionMountain	0.055	-0.122
geo_regionNE Central	0.021	-0.156
geo_regionNW Central	0.180	0.002
geo_regionPacific	0.115	-0.061
geo_regionSE Central	0.254	NA
geo_regionSouth Atlantic	0.141	-0.039
geo_regionSW Central	0.190	0.009
moresp_natforeignaid1	0.103	0.100
moresp_natarms1	-0.181	-0.178
moresp_natcrime1	-0.025	-0.025
moresp_natdrug1	0.168	0.170
moresp_nateduc1	0.181	0.182
moresp_natenvir1	0.130	0.128
moresp_natwelfare1	0.297	0.298
moresp_nathealth1	0.306	0.305
moresp_natrace1	0.285	0.282
moresp_natspace1	-0.030	-0.030
pol_liberalism	0.417	0.416
race_textBlack	0.000	0.000
geo_regionMiddle Atlantic	NA	-0.205

Part 3: Model Validation

In this third stage, I compared the candidate models by testing them using the validation set (25% of the full sample). To clarify, I mean that I fit each of the candidate models using the full model-building set, and then tested them against the validation set. Below, we can see the test MSE by model.

The results below indicate that the best candidate model is associated with the PCR: i.e., the PCR with 40 principal components. My OLS model was ranked in 6th place. However, it's important to note that the test MSEs are again quite similar across the candidate models (and certainly among the top 8 performers). Since the MSEs are so similar, it's reasonable to expect that the rank order could change if the validation set had been slightly different.

Because of this reason, I've decided to select my OLS model as the finalist. Although it performed slightly

less well than the 1st place method/model, the OLS is preferable because unlike the PCR, it actually generates regression coefficients – which are relatively easy to interpret. In addition to the benefits of high interpretability, selecting the OLS also means that I can test the interaction between political liberalism and support for the Democratic Party – which, as discussed earlier, is interesting for theoretical reasons.

rank	method	test_mse
1	pcr_40m	3.013
2	lasso_min	3.021
3	ridge_min	3.026
4	fwd_selection	3.045
5	back_selection	3.049
6	ols	3.052
7	lasso_1se	3.080
8	ridge_1se	3.089
9	pls_4m	3.233

Part 4: Model Testing

In the final stage, I will take my final model (which was fit using the model-building set) and test it using the held-out test set. When I do this, my test MSE which is 3.005; this is actually lower than the lowest test MSE achieved during the previous validation stage – in which the PCR model with 40 principal components achieved the lowest test MSE of 3.013. This supports the argument that although the OLS did not perform the best during the validation stage, it is still a good model relative to its competitors.

model	test_MSE
best_OLS	3.005

Conclusion

In my conclusion, I want to address a few implications of my findings. To recap, the two **research questions** I posed at the start of this report were the following:

(1) What explains support for political parties in the U.S. context? The focus here is on identifying statistically and substantively significant predictors (i.e., which specific predictors matter? And how much do these individual predictors matter?)

(2) Overall (i.e., multiple relevant predictors are used), how well can we predict support for political parties? How good is the "best" model? The focus here is on assessing the adjusted R-squared and the size of the test MSE.

First, it's clear that there are many statistically significant predictors of support for the Democrats (DV). We know this because the automated variable selection methods (e.g., lasso, forward selection) kept many predictors in the model. However, without looking at the relative size of the actual regression coefficients themselves, it's not necessarily clear whether the predictors matter in a substantive sense. For example, it's possible for a predictor to have a very small p-value (e.g., less than 0.00001) even though the substantive effect is small (e.g., perhaps a one-unit increase in the X results in less than a 1% change in the Y value). This would occur if the variance in X is very small; that is, the X values are tightly clustered around the mean value.

To directly address the issue of substantive significance, I've displayed the results of my final OLS model below. The results suggest that there are indeed a number of substantively significant predictors.

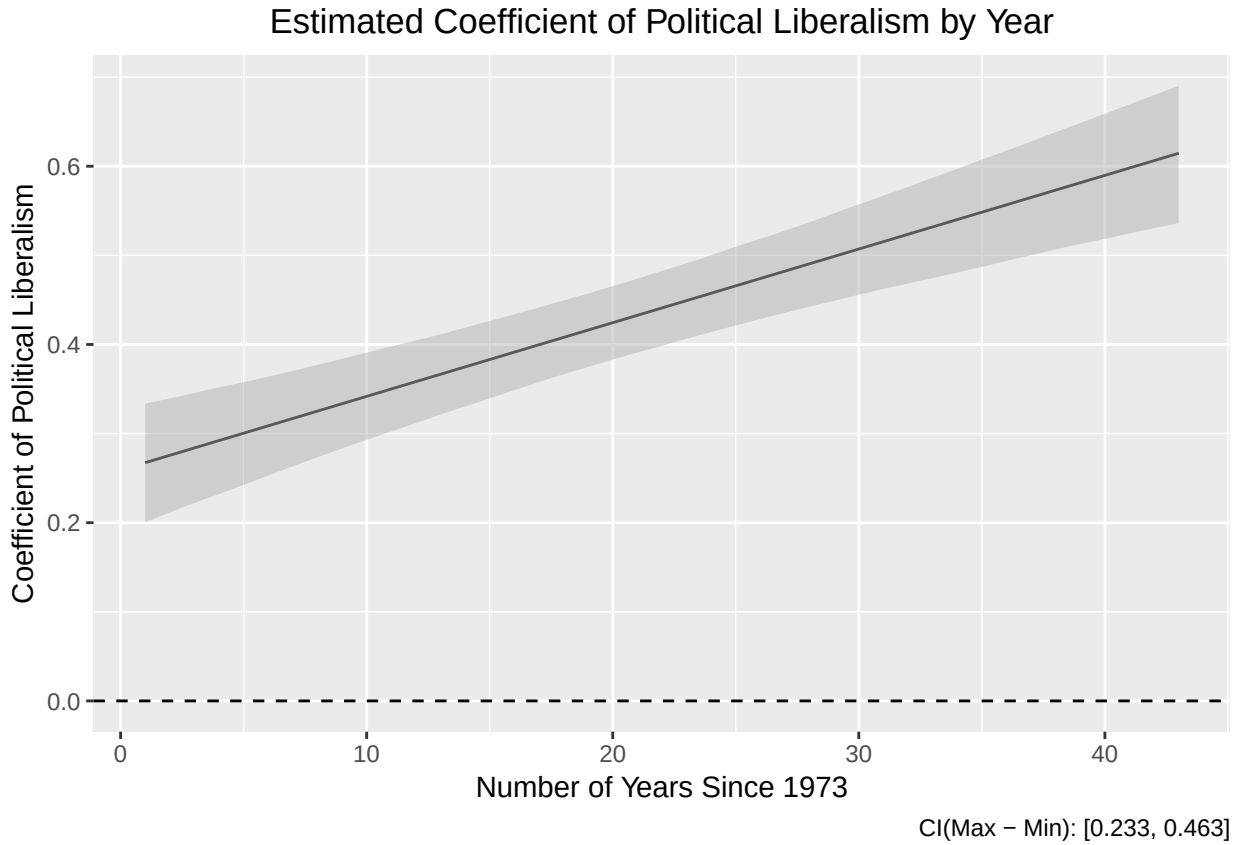
For example, after accounting for the other predictors in the model, black respondents have a mean level of support for the Democrats that is about 1.2 points higher than that of whites – this is a substantial difference, given that the mean level of support for Democrats is about 4.2 on a 1-7 point scale. Relative to those who self-identify as middle class, those in the upper class have a mean level of support that is about .64 points lower, which is more than 10% of the mean value of the DV.

Other statistically and substantively significant predictors of support for the Democrats includes support for increasing public spending on welfare, drug problems, defense; religion, and gender.

term	estimate	std.error	statistic	p.value
(Intercept)	3.316	0.371	8.927	0.000
year	-0.042	0.006	-6.948	0.000
age	0.009	0.002	5.173	0.000
race_textBlack	1.207	0.102	11.887	0.000
race_textOther	0.212	0.143	1.481	0.139
female1	0.130	0.053	2.453	0.014
religionCatholic	0.131	0.097	1.355	0.176
religionChristian	-0.229	0.279	-0.820	0.412
religionJewish	0.651	0.187	3.480	0.001
religionOrthodox-Christian	-0.116	0.782	-0.148	0.882
religionProtestant	-0.332	0.091	-3.643	0.000
subj_classid_textLower Class	0.057	0.138	0.410	0.682
subj_classid_textOther	-0.281	0.552	-0.509	0.611
subj_classid_textUpper class	-0.636	0.149	-4.281	0.000
subj_classid_textWorking Class	0.251	0.059	4.291	0.000
ln_famincome	-0.038	0.030	-1.242	0.214
yrs_educ	-0.031	0.011	-2.959	0.003
pol_liberalism	0.260	0.035	7.447	0.000
moresp_natwelfare1	0.359	0.073	4.919	0.000
moresp_natdrug1	0.214	0.055	3.905	0.000
moresp_natarms1	-0.146	0.063	-2.309	0.021
moresp_natenvir1	0.188	0.057	3.297	0.001
moresp_natrace1	0.327	0.064	5.108	0.000
year:pol_liberalism	0.008	0.001	5.758	0.000

In addition, the OLS model above also indicates that the interaction term `year:pol_liberalism` is statistically significant. This means that the strength of the association between political liberalism (or having a liberal ideology) and support for the Democrats is shaped by year. In particular, the interaction term has a positive coefficient, which implies that the link between ideology and party ID has become stronger over time. To help readers visualize this relationship, I've created a graph that shows how the size of the coefficient for political liberalism changed between 1974 and 2016.

In particular, we can see that the coefficient increases from about 0.27-0.28 to about 0.62 between 1974-2016. In other words, the strength of the association between ideology and party membership doubles in a quantitative sense. In the early 1970s, the mean difference in support for the Democrats among ideological moderates and those who were very liberal was about 0.84 points (or `3*.28`). By 2016, the difference in support for the Democrats among those who were moderate v. very liberal increased to over 1.86 (`3*.62`). This is quite significant, given that the SD of the DV is about 2 points.



Second, how well can the models predict support for the Democrats? It's important to note that while this is related to the first question, it is still distinct: the second question is by definition concerned about the overall accuracy of the predictions given the full set of predictors in a given model. This question is less concerned with the predictive ability of any individual predictor.

If we can recall, the candidate models achieved test MSEs of about 3. The square root of the MSE, or RMSE, is a measure of the average difference between the predicted and actual DV values. To determine whether an RMSE is "good", we need a baseline or point of reference. We can use the standard deviation (SD) of the dependent variable: it represents the average deviation of the DV values from their mean value. In this case, the RMSE is about 1.7 and the SD of the DV (or Y) is around 2.0. This means that on average, the difference between the predicted and actual values of the DV is about as large as the average deviation of the DV from the mean value – so unfortunately, the model is not doing that well!

The adjusted R-squared for this model is about .23, which means that the model is only able to explain about 23% of the variation in the DV; this largely supports the discussion above. In a future iteration of this project, I may try to increase my predictive ability by looking for new data and/or using more complex modeling strategies. For example, I may try fitting a RE model that accounts for the hierarchical structure of the data (e.g., individuals are nested within regions).