Report 3

Wei Zhong

4/3/2017

Introduction

As per the Constitution, the U.S. House of Representatives makes and passes federal laws. The House is one of Congress's two chambers (the other is the U.S. Senate), and part of the federal government's legislative branch. The number of voting representatives in the House is fixed by law at no more than 435, proportionally representing the population of the 50 states.[1] The House of Representatives are elected to a two-year term, each representative serves the people of a specific congressional district by introducing bills and serving on committees, among other duties.

Congressional districts are the 435 areas from which members are elected to the U.S. House of Representatives. After the apportionment of congressional seats among the states, which is based on decennial census population counts, each state with multiple seats is responsible for establishing congressional districts for the purpose of electing representatives. Each congressional district is to be as equal in population to all other congressional districts in a state as practicable. The boundaries and numbers shown for the congressional districts are those specified in the state laws or court orders establishing the districts within each state.

Congressional districts for the 108th through 112th sessions were established by the states based on the result of the 2000 Census. Congressional districts for the 113th through 115th sessions were established by the states based on the result of the 2010 Census. Boundaries are effective until January of odd number years (for example, January 2015, January 2017, etc.), unless a state initiative or court ordered redistricting requires a change. All states established new congressional districts in 2011-2012, with the exception of the seven single member states (Alaska, Delaware, Montana, North Dakota, South Dakota, Vermont, and Wyoming).[2]

The 113th United States Congress was a meeting of the legislative branch of the United States federal government, from January 3, 2013, to January 3, 2015. It was composed of the United States Senate and the United States House of Representatives based on the results of the 2012 Senate elections and the 2012 House elections. The seats in the House were apportioned based on the 2010 United States Census. It first met in Washington, D.C. on January 3, 2013, and it ended on January 3, 2015. Senators elected to regular terms in 2008 were in the last two years of those terms during this Congress.[3]

The modern political party system in the U.S. is a two-party system dominated by the Democratic Party and the Republican Party. These two parties have won every United States presidential election since 1852 and have controlled the United States Congress to some extent since at least 1856.[4]

The Democratic Party is one of two major political parties in the U.S. Founded in 1828 by Andrew Jackson, it is the oldest extant voter-based political party in the world. It since 1912 has positioned itself as the liberal party on domestic issues. The economic philosophy of Franklin D. Roosevelt, which has strongly influenced modern American liberalism, has shaped much of the party's agenda since 1932. Roosevelt's New Deal coalition controlled the White House until 1968 with the exception of Eisenhower 1953–1961. Since the mid-20th century, Democrats have generally been in the center-left and currently support social justice, social liberalism, a mixed economy, and the welfare state, although Bill Clinton and other New Democrats have pushed for free trade and neoliberalism, which is seen to have shifted the party rightwards. Democrats are currently strongest on the East and West Coasts and in major American urban centers. African-Americans and Latinos tend to be disproportionately Democratic, as do trade unions.[5]

[1] http://www.house.gov/content/learn/

[2] https://www.census.gov/geo/maps-data/data/aboutcd.html

[3] https://en.wikipedia.org/wiki/113th_United_States_Congress

[4] https://en.wikipedia.org/wiki/Political_parties_in_the_United_States

[5] https://en.wikipedia.org/wiki/Political_parties_in_the_United_States

The Republican Party is another major contemporary political party. Since the 1880s it has been nicknamed (by the media) the "Grand Old Party" or GOP, although it is younger than the Democratic Party. Since its founding, the Republican Party has been the more market-oriented of the two American political parties, often favoring policies that aid American business interests. As a party whose power was once based on the voting clout of Union Army veterans, this party has traditionally supported more aggressive defense measures and more lavish veteran's benefits. Though initially founded to oppose slavery, following Richard Nixon's "Southern Strategy" in 1968, the Republican Party has become the less progressive party in areas of racial, gender and identity politics-motivated social justice. Today, the Republican Party supports an American conservative platform, with further foundations in economic liberalism, fiscal conservatism, and social conservatism. The Republican Party tends to be strongest in the Southern United States and the "flyover states", as well as suburban and rural areas in other states. One significant base of support for the Republican Party are Evangelical Christians, who have wielded significant clout in the party since the early 1970s.[6]

Kramer (1971) emphasizes that voters are heterogenous and disaggregate electorate into three informational groups based upon education attainment. Mutz and Mondak (1997) note that, "in studies of American political behavior it is axiomatic that groups matter" (Leighley, 2010; 386). These groups can be created from any set of criteria, from race to religion to special interests. As shown by Hibbs, et al. (1982), these groups view elections heterogeneously. This means that each group will have a different preference for which party they choose to vote for. Seltzer and Hutto (2013) supported this idea, finding a racial difference in the perception of how the U.S. national economy was doing depending on who was the president at that time. Specifically, they observed a racial difference before the 2008 election, where whites were more likely than blacks to say that the national economy was doing, which switched once Obama was elected (Seltzer and Hutto, 2013). Ansolabehere et. al. (2014) also observed evidence that groups evaluate the economy heterogeneously. They discover that "individuals from groups that experience more unemployment report the national unemployment rate is higher" (Ansolabehere et. al., 2014; 381). Breaking this finding down into more specific groups, they note that ethnic minorities with lower educational attainment and individuals from states with higher unemployment rates perceive that there is a higher rate of national unemployment (Ansolabehere et. al., 2014; 381). In short, literature recognizes individual heterogeineity in shaping political choice and shows that socio-demographical factors (such as income, social status, employment, ethnicity, age, gender, education) drive people's choice toward different parties, given distinct party label (as mentioned above, Democratic party tends to promote social justice, social liberalism, a mixed economy, and the welfare state, while Republicans incline to be more market-oriented, often favoring policies that aid American business interests. Economically, Democratic party focuses on unemployment, while Republicans emphasize reducing inflation).

Data

There are total 436 congressional districts including the federal district of Washington, D.C.. The data is consist of 14 variables: three of them are categorical, and the rest are numeric. The statistical summary of each numerical variable are shown in Table 1. Among them, the total population variable has the largest standard variation, with a value of 34305, which implies a substantially variability of population among districts. The most populous congressional district is in Montana, which has a value of 998199, while the least populous district is in Rhode Island, with a total population of 524097. Table 2 displays the summary characteristics of categorical variables including states, districts, and party choice. Specifically, Table 2 shows all 50 states and the federal district of Washington, D.C., and the total number of congressional districts for each state. Within each state, the column 2 and 3 give the number of districts voting for either Democratic or Republican party in the 113th Congress. In addition, at aggregate level, there are 202 districts voting for Democrats, and 234 districts voting for Republicans.

[6] https://en.wikipedia.org/wiki/Political_parties_in_the_United_States

Table 1: Statistical summary of socio-demographical variables

Statistic	N	Mean	St. Dev.	Min	Max
Total Population	436	714,660.000	34,305.000	524,097	998,199
Median Age	436	37.490	3.607	27.300	51.900
Medina HouseIncome	436	52,205.000	13,874.000	23,894	109,505
Median FamilyIncome	436	63,381.000	16,799.000	26,695	127,939
Percent Unemployed	436	10.380	2.938	3.100	24.700
Percent of people withincome below poverty	436	15.970	5.816	4.100	39.400
Percent high school graduates ($\geq$ 25 age)	436	85.710	6.728	50.900	95.900
Percent with bachelor's degree ($\geq$ 25 age)	436	28.280	10.060	8.200	68.600
Percent Black or African-American	436	13.620	14.510	0.700	65.600
Percent Hispanic or Latino	436	16.680	17.940	0.700	86.600

Predicting Party Choice: Model Building

In this section, I fit binary regression models with "party" as the response. In order to make sure that all models are for the probability of Democrat, I create another new variable called "dem", which is coded as 1 if Democratic party, and 0 if Republican party. In order to examine the effects of socio-demographical factors on party choice, I fit four different models: a main effect only logistic regression; a logistic regression with all main effects and also all two-way interactions; a main-effects-only probit regression with same variables; and a probit regression with all main effects and also all two-way interactions. All models are fitted based on same nine variables including medAge, medHouseIncome, medFamilyIncome, pctUnemp, pctPov, pctHS, pctBach, pctBlack, and pctHisp. For each model, I apply backward elimination. Table 3 shows the AIC values of all four of the resulting models. Applying backward elimination, the main-effect and two way interactions logistical model has the lowest AIC, with a value of 402.6. Among those nine variables, median household income and the percent high school graduates are removed in both main-effect-only logit model and main-effect-only probit model.

Predicting Party Choice: Model Assessment

In the previous section, the logistic regression with all main effects and also all two-way interactions give the lowest AIC value. Consequently, I choose this model as the best model for predicting party choice according to AIC criterion. The correlation measure of the model can be calculated as:

$$R(y, \hat{\mu}) = 0.7087$$

In order to assess the model, I use a classfication table and summarize the predictive power by:

$$\text{sensitivity} = P(\hat{y} = 1|y = 1) = 0.797, \quad \text{specificity} = P(\hat{y} = 0|y = 0) = 0.8718$$

In addition, I try a "leave-one-out" cross validation classification method which classify each observation using a model fit without it. The corresponding estimated sensitivity and specificity are given as below:

$$\text{sensitivity} = P(\hat{y} = 1|y = 1) = 0.7178, \quad \text{specificity} = P(\hat{y} = 0|y = 0) = 0.8205$$

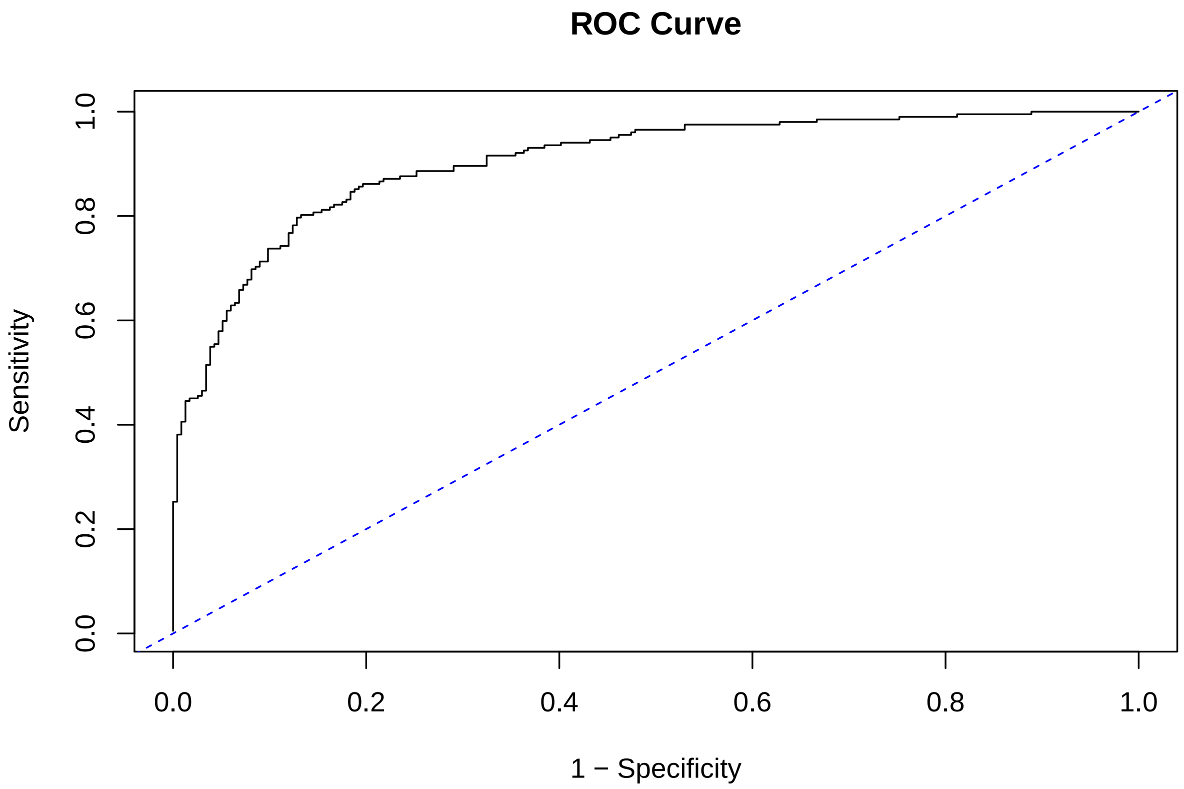
Futhermore, a receiver operating characteristic (ROC) curve can be more informative than a classification table, since it summarizes predictive power for all possibile π_0. For this case, a ROC curve is displayed as following. And the area under the ROC curve is 0.9044. In a summary sense, the greater the area under the ROC curve, the better the predictions. In other words, the model has a pretty good predictive power of party choice.

Table 2: Summary of congressional districts and its party choice for each state

	# dists voting for Dem	# dists voting for Rep	Total# dists
Alabama	1	6	7
Alaska	0	1	1
Arizona	5	4	9
Arkansas	0	4	4
California	38	15	53
Colorado	3	4	7
Connecticut	5	0	5
Delaware	1	0	1
District of Columbia	1	0	1
Florida	10	17	27
Georgia	5	9	14
Hawaii	2	0	2
Idaho	0	2	2
Illinois	12	6	18
Indiana	2	7	9
Iowa	2	2	4
Kansas	0	4	4
Kentucky	1	5	6
Louisiana	1	5	6
Maine	2	0	2
Maryland	7	1	8
Massachusetts	9	0	9
Michigan	5	9	14
Minnesota	5	3	8
Mississippi	1	3	4
Missouri	2	6	8
Montana	0	1	1
Nebraska	0	3	3
Nevada	2	2	4
New Hampshire	2	0	2
New Jersey	6	6	12
New Mexico	2	1	3
New York	21	6	27
North Carolina	4	9	13
North Dakota	0	1	1
Ohio	4	12	16
Oklahoma	0	5	5
Oregon	4	1	5
Pennsylvania	5	13	18
Rhode Island	2	0	2
South Carolina	1	6	7
South Dakota	0	1	1
Tennessee	2	7	9
Texas	12	24	36
Utah	1	3	4
Vermont	1	0	1
Virginia	3	8	11
Washington	6	4	10
West Virginia	1	2	3
Wisconsin	3	5	8
Wyoming	0	1	1

Table 3: AIC values for the four different models

	AIC
Main-effect-only logit	456.8
Main-effect and two-way interactions logit	402.6
Main-effect-only probit	459.3
Main-effect and two-way interactions profit	404.3



Conditional Dependence of Party and Affluence

In this section, I create a new data set by removing all state and (the DC) having only one district. I also create a wealth indicator variable that designates districts with median household income exceeding \$52000 as "Wealthy" and others as "Non-Wealthy". If the median household income is wealthy, coded as 1, otherwise, 0. To examine the conditional dependence of party and wealth, I fit a logistic regression based on the variables wealth and state. The corresponding results are shown in Table 4. Consquently, the conditional odds ratio for electing a Democrat for Wealthy disctrict versus Non-wealthy district can be written as:

$$\exp(-1.099) = 0.3329$$

And an approximate 95% confidence interval for the odds ratio is given by:

$$\exp(-1.099 \pm 1.96 \times 0.294) = (0.1871, 0.5923)$$

The conditional odds ratio 33.29%, which is smaller than 1, means that wealthy district are less likely to electing Democrats compared to non-wealthy district. In addition, the approximate 95% confidence interval for the odds ratio (0.1871, 0.5923) does not include 1, further validating that wealthy district are less likely to electing Democrats compared to non-wealthy district.

Alternatively, the logistic regression without stratification by state, but still using only the districts used in the analyses give a marignal odds ratio, with a value of $\exp(0.213) = 1.273$. In other words, the marginal odds ratio give a conflicting result against the conditional odds ratio. In the case without stratification, wealthy districts seems more likely to elect Democratic party than non-wealthy district. The possible explanation for this conflicting result might be stability of state historical election results, that is red states and blue states. Since 2000, the political positions of the states are consistent in the short term from year to year; for example, New York has strongly favored the Democrats in recent decades, Utah has been consistently Republican, and Ohio has been in the middle. The existing literature in political science shows strong correlations in the political party share of the vote in each state from one presidential election to the next. Therefore, for predicting party choice, we should consider the state's relative partisanship.

Furthermore, I give a Mantel-Haenszel analysis, which proposes a non-model-based test of H_0: conditional independence in $2 \times 2 \times K$ tables. In our case, I create a $2 \times 2 \times 43$ tables by removing the states and the DC which has only one district. The resulting common odds ratio is 0.3602, and a 95% confidence interval for the common odds ratio is (0.2054, 0.6315), which means the odds of electing democracy may be as much as 80% lower with the wealthy district, or they may be little as 36% as low.

Conclusion

Given the analysis above, the party choice in the district levels can be predicted by the socio-demographical variables such as median age, meidan household income, median family income, unemployment rate, poverty rate, education, and ethnicity. Furthermore, very different political party economic policies lead voters to make distinct choices. Through the analysis, by controlling state effect, we could conclude that wealthy districts tend to favor Republicans, while non-wealthy districts are more likely to support Democratic party. In other words, voters' party choice is substantially correlated with parties' label and policies.

Table 4: The logistical regression result of party choice and wealth, stratified by states

	Dependent variable:
	Democratic Party
wealth	-1.099^{***} (0.294)
stateArizona	2.278^{*} (1.291)
stateArkansas	-15.880 (1,978.000)
stateCalifornia	3.339^{***} (1.146)
stateColorado	2.170 (1.354)
stateConnecticut	20.350 (1,769.000)
stateFlorida	1.184 (1.157)
stateGeorgia	1.293 (1.225)
stateHawaii	20.350 (2,797.000)
stateIdaho	-15.880 (2,797.000)
stateIllinois	3.037^{**} (1.214)
stateIndiana	0.519 (1.352)
stateIowa	1.946 (1.495)
stateKansas	-15.650 (1,928.000)
stateKentucky	0.072 (1.542)
stateLouisiana	0.204 (1.547)
stateMaine	19.870 (2,695.000)
stateMaryland	4.631^{***} (1.548)
stateMassachusetts	20.140 (1,288.000)
stateMichigan	1.293 (1.225)
stateMinnesota	2.920^{**} (1.334)
stateMississippi	0.583 (1.585)
stateMissouri	0.687 (1.362)
stateNebraska	-15.570 (2,214.000)
stateNevada	1.946 (1.495)
stateNew Hampshire	20.350 (2,797.000)
stateNew Jersey	2.607^{**} (1.257)
stateNew Mexico	2.374 (1.637)
stateNew York	3.627^{***} (1.199)
stateNorth Carolina	1.006 (1.244)
stateOhio	0.800 (1.234)
stateOklahoma	-15.880 (1,769.000)
stateOregon	3.344^{**} (1.577)
statePennsylvania	1.146 (1.215)
stateRhode Island	19.870 (2,695.000)
stateSouth Carolina	0.000 (1.535)
stateTennessee	0.428 (1.350)
stateTexas	1.364 (1.147)
stateUtah	1.354 (1.617)
stateVirginia	1.339 (1.297)
stateWashington	2.780^{**} (1.286)
stateWest Virginia	0.988 (1.637)
stateWisconsin	1.542 (1.322)
Constant	-1.681 (1.086)
Observations	428
Log Likelihood	-223.900
Akaike Inf. Crit.	535.800
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix

```r
rm(list = ls())
getwd()
setwd("/Users/Rocio/Library/Mobile Documents/com~apple~CloudDocs/STAT_426")
dta  = read.csv("party113cong.csv", header = T, sep = ",")
str(dta)

summary(dta[, 4:13])
library(stargazer)
stargazer(dta[, 4:13], digits = 1)

# sort the data according to the total population
df.pop = dta[order(dta$totPop),]

# summary of categorical vars: state, party choice and districts
df3 = xtabs(~state + party, data=dta)
df4 = as.data.frame.matrix(df3)
xtable(df4)

# fit main effect model
dta$dem = ifelse(dta$party == "D", 1, 0)
fit1 = glm(dem ~ medAge+ medHouseIncome + medFamilyIncome +
             pctUnemp + pctPov + pctHS + pctBach + pctBlack +
             pctHisp, family = binomial, data = dta)

summary(fit1)

# main effect and two way interactions
fullmod1  = glm(dem ~ (medAge+ medHouseIncome + medFamilyIncome +
                 pctUnemp + pctPov + pctHS + pctBach + pctBlack + pctHisp)^2,
                family = binomial, data = dta)

# main-effects-only probit
fit2 = glm(dem ~ medAge+ medHouseIncome + medFamilyIncome +
             pctUnemp + pctPov + pctHS + pctBach + pctBlack + pctHisp,
           family = binomial(link = "probit"), data = dta)

# main effect and two way interactions probit
fullmod2  = glm(dem ~ (medAge + medHouseIncome + medFamilyIncome +
                         pctUnemp + pctPov + pctHS + pctBach + pctBlack + pctHisp)^2,
                family = binomial(link = "probit"), data = dta)

# backward elimination for main logit model
backfit1 = step(fit1)
summary(backfit1) # houseincome, popHS removed

# backward elimination for full logistic model
backmod1 = step(fullmod1)
summary(backmod1) ## best model, smallest AIC

# backward elimination for main probit model
backfit2 = step(fit2)
```

```
summary(backfit2) # houseincome, pctHS removed

# backward elimination for full probit model
backmod2 = step(fullmod2)
summary(backmod2)

# correlation measure
cor(dta$dem, fitted(backmod1))

# an apparent classification table, with estimated sensitivity and specificity

pi0 = 0.5
table(y=dta$dem, yhat=as.numeric(fitted(backmod1) > pi0))

# apparent sensitivity
161/(41 + 161)

# apparent specificity
204/(204 + 30)

# a cross-validated classification table, with estimated sensitivity and specificity
pihatcv <- numeric(nrow(dta))

for(i in 1:nrow(dta))
  pihatcv[i] <- predict(update(backmod1, subset=-i), newdata=dta[i,],
                        type="response")

table(y=dta$dem, yhat=as.numeric(pihatcv > pi0))

# cross-validated sensitivity
145/(57 + 145)

# cross-validated specificity
192/(192 + 42)


# ROC Curve (Model 3)

n <- nrow(dta)
pihat <- fitted(backmod1)

true.pos <- cumsum(dta$dem[order(pihat, decreasing=TRUE)])
false.pos <- 1:n - true.pos

plot(false.pos/false.pos[n], true.pos/true.pos[n], type="l",
     main="ROC Curve", xlab="1 - Specificity", ylab="Sensitivity")
abline(a=0, b=1, lty=2, col="blue")


mean(outer(pihat[dta$dem==1], pihat[dta$dem==0], ">") +
       0.5 * outer(pihat[dta$dem==1], pihat[dta$dem==0], "=="))
# area under curve (concordance index)
# Graduate section
# Conditional Dependence of Party and Affluence
# remove all state which has 1 district and DC
library(plyr)
newdf = as.data.frame( dta %>% group_by(state) %>% filter(n() > 1))

# just for check
df  = count(newdf, 'state')
which(df$freq == 1)

# create a wealth indicator var.
# districts with median household income exceeding $52000 as "Wealthy" and others as "Non-Wealthy"
which(newdf$medHouseIncome == 52000)
newdf$wealth = ifelse(newdf$medHouseIncome > 52000, 1, 0) # wealty = 1, non-wealthy = 0

fitwealth = glm(dem ~ wealth + state, family = binomial, data = newdf)
summary(fitwealth)

exp( -1.10e+00)  # MLE of common conditional odds ratio
exp( -1.10e+00  + c(-1,1) * 1.96 * 2.94e-01)  # transformed Wald interval

# marginal odds ratio
fitmar = glm(dem ~ wealth, family = binomial, data = newdf)
summary(fitmar)

# Cochran-Mantel-Haenszel Approach
newdf.array <- xtabs(~dem + wealth + state, data= newdf, drop.unused.levels = T)
newdf.array

mantelhaen.test(newdf.array, correct=FALSE)
```