1. Representatives of the House represent – and are elected by – residents of their corresponding geographical area that is known as a "congressional district." Congressional districts are allocated to each of the 50 states on a basis of population as measured by the U.S. Census, with each district entitled one representative. There are 435 congressional districts, hence 435 Representatives. In addition to the 50 U.S. states, Washington D.C. also has a representative in the house, who cannot vote on the floor but can vote on procedural matters[1].

Both the House of Representatives and the Senate are re-elected every two years. Each duration of two years is numbered since the independence of the United States. The data we have for this report is on the 113th Congress, which was in meeting from January 3, 2013 to January 3, 2015.

One of the most important factors for Representatives' election is the political party they belong to. Theoretically, a Representative can come from any political party. However, given the two-party political structure of the United States, it is not surprising to know that there is seldom a House Representative who comes from any other party than the Republican or the Democratic. This is also the case in the 113th Congress's House, where each representative is either Republican or Democrat [2].

Anecdotal observation seems to show party preference is related to some demographic characteristics. For example, there is a commonly held notion that as an individual grows older, he or she becomes more conservative, thus may identify more with the Republican Party. Another usually observed phenomenon is the Democratic Party's identity politics may appeal to non-whites, thus congressional districts with more non-whites may prefer the Democratic Party to the Republican. However, we could not draw any scientific conclusion from merely anecdotal evidence. Therefore, it might be of interest to relate the party of an elected Representative to demographic characteristics of his or her district with statistical methods.

1. Data

The dataset we have contains information on congressional districts' demographic characteristics and party representation. There are 436 observations, with each observation being one congressional district, as well as Washington D.C. The dataset has 13 columns: the first column is unique identifiers for each district, which will not be included in analysis. Among the remaining 12 variables, 10 are numeric variables and 2 are categorical.

Information about numerical variables is listed in the table below.

[1] https://en.wikipedia.org/wiki/District_of_Columbia_voting_rights
[2] https://en.wikipedia.org/wiki/113th_United_States_Congress

#	Variable Name	Description	Minimum	Mean	Maximum
1	totPop	total population	524,097	714,660	998,199
2	medAge	median age	27.3	37.5	51.9
3	medHouseIncome	median house income	23,894	52,205	109,505
4	medFamilyIncome	median family income	26,695	63,381	127,939
5	pctUnemp	percent unemployed	3.1	10.4	24.7
6	pctPov	percent below poverty line	4.1	16.0	39.4
7	pctHS	percent high school graduates	50.9	85.7	95.9
8	pctBach	percent with bachelor's degree	8.2	28.3	68.6
9	pctBlack	percent African-American	0.7	13.6	65.6
10	pctHisp	percent Latino	0.7	16.7	86.6

Total population in each congressional district demonstrates a considerable amount of variability, with the largest population almost twice as big as the smallest one. Analysis shows that the standard deviation of totPop is $34,305$.

Two categorical variables are state and party. In the state variable, California has the largest frequency, with 53 congressional districts. Smallest frequency is 1, which occurs to several states, namely AL, DC, MT, ND, SD, VT and WY. On party affiliation, 234 are Republicans, and 202 are Democrats. There is no independent.

2. Predicting Party Choices: Model Building

As is instructed, I fit four binary regression models with party as response based on variables medAge, medHouseIncome, medFamilyIncome, pctUnemp, pctPov, pctHS, pctBath, pctBlack, and pctHisp. All four models are fit for the probability of D (Democrat. The dataset was modified accordingly). The four models are:

- fullmod1: a main-effects-only logistic regression;
- fullmod2: a logistic regression with all main effects and also all two-way interactions – but not squared terms;
- fullmod3: a main-effects-only probit regression;
- fullmod4: a probit regression with all main effects and also all two-way interactions – but not squared terms;

Then, for each of the four models, I applied backward elimination and obtained an AIC-optimal model from each of them. The table below displays information on the four models after backward elimination.

#	Model base	AIC	Variables Eliminated
1	Main-effect Logit	456.84	medHouseIncome, pctHS
2	Two-way Interaction Logit	402.62	None
3	Main-effect Probit	459.34	medHouseIncome, pctHS
4	Two-way Interaction Probit	404.25	None

Backward elimination eliminates `medHouseIncome` and `pctHS` from each of the two main-effect only models. No variable is completely eliminated from the two two-way interaction models.

3. Predicting Party Choice: Model Assessment

In the previous section, two-way interaction logistic model achieved the lowest AIC after backward elimination. Analysis in this section is based on that model.

$R(\mathbf{y}, \hat{\mu})$ is the correlation between the observed response and the model's fitted value. In logistic regression with ungrouped data, it is the correlation between the n binary y_i observations and the estimated probability. For the chosen model, $R(\mathbf{y}, \hat{vec\mu}) = 0.7087$.

Classification table is given below:

	Predicted Republican	Predicted Democrat
Actual Republican	204	30
Actual Democrat	41	161

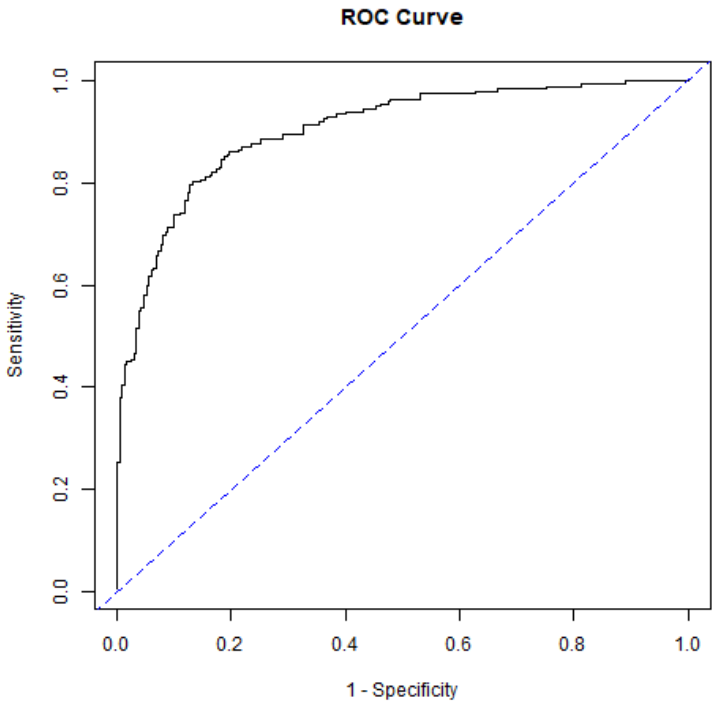
Sensitivity is 79.70%. Specificity is 87.18%.

Here is a cross-validated classification table, with estimated sensitivity and specificity.

	Predicted Republican	Predicted Democrat
Actual Republican	192	42
Actual Democrat	57	145

Sensitivity is 71.78%. Specificity is 82.05%.

ROC curve is graphed below. In this ROC curve, concordance index is calculated to be 0.9044.



4. Conditional Dependence of Party and Affluence

In this section, I will present data analysis result based on states with at least two congressional districts. I removed all states with only one congressional district, including Washington D.C., from the dataset.

Another variable, labeled `wealth`, was created based on median household income, with congressional districts having `(medHouseIncome)` exceeding $52,000 labeled as Wealthy, and others labeled as Non-Wealthy. In the regression models in this section, I test the dependence of party preference on wealth, conditional on states.

4.1 Logistic Regression

Logistic regression model is given by

$$\text{logit}\pi_D = \alpha + \beta_1 * \texttt{wealth} + \beta_{\texttt{state}}$$

In this formula, $\beta_{\texttt{state}}$ takes a value depending on which state it represents. β_1 is the conditional odds ratio that we are interested in.

Logistic regression returns $\hat{\beta}_1 = -1.099$. Therefore, the conditional odds ratio for electing a Democrat for Wealthy district versus Non-Wealthy district is $e^{-1.099} = 0.3332$ .The standard error is 0.2943. We can calculate the 95% confidence interval of $\hat{\beta}_1$ to be

$$-1.099 \pm 1.96 * \sqrt{0.2943} = [-2.162, -0.03571]$$

, and

$$e^{[-2.162,-0.03571]} = [0.1150, 0.9649]$$

which indicates that we are 95% confident that given a specific state, the odds ratio to elect a Democrat representative for a wealthy congressional district versus a non-wealthy district is between 0.1150 and 0.9649.

If we fit a logistic regression model to predict `party` on `wealth` only, we get logged odds ratio to be 0.2130, indicating wealthy districts are more likely to elect a Democrat, which contradicts the conclusion that we draw from the conditional odds model. This apparent discrepancy might be able to be explained by the difference between the in-state political opinion spectrum and national political opinion spectrum: Within a specific state, wealthier congressional district might be more inclined towards the Republican. But nationwide, wealthier states are also states that are more liberal (such as the West coast, the East coast), resulting in an apparent positive correlation between preference of Democrats to Republicans and affluence.

4.2 Mantel-Haenszel Analysis

Mantel-Haenzsel analysis returns the odds ratio for electing a Democrat for Wealthy districts versus non-wealthy districts is 0.3602, with a 95% confidence interval being [0.2054, 0.6315]. This result demonstrates that conditional upon states, wealthy districts are less likely to elect a Democrat than non-wealthy districts. The result is consistent with the conclusion we obtained from the model-bases conditional odds test.

5. Conclusions

From the analysis above, we could draw a few conclusions regarding the relationship between party preference and demographic characteristics of congressional districts. First, a congressional district's party preference is associated with its population's age, income, education, and racial composition; Second, conditional on states, a wealthy congressional district is less likely to elect a Democrat than a non-wealthy district. Third, on a national scale, a wealthy congressional district is more likely to elect a Democrat than a non-wealthy district.

6. Appendix

```
# Part1

# Reading data
party113cong <- read.csv('D:/Zhao/Documents/Spring_2017/426/
party113cong.csv', header = TRUE)
```

```
head(party113cong)


# Part 2

# Descriptive Statistics

apply(party113cong[, c(-1, -2, -3)], 2, min)
apply(party113cong[, c(-1, -2, -3)], 2, mean)
apply(party113cong[, c(-1, -2, -3)], 2, max)

sd(party113cong$totPop)

sort(summary(party113cong$state))

summary(party113cong$party)

## Part 3

# Make sure 'R' is the reference level:

party113cong$party <- relevel(party113cong$party, ref = "R")

## Fitting Main-effect only logistic model

fullmod1 <- glm(party ~ medAge + medHouseIncome + medFamilyIncome +
pctUnemp + pctPov + pctHS + pctBach + pctBlack + pctHisp,
family = binomial, data = party113cong)

fullmod2 <- glm(party ~ (medAge + medHouseIncome + medFamilyIncome +
 pctUnemp + pctPov + pctHS + pctBach + pctBlack + pctHisp) ^ 2,
 family = binomial, data = party113cong)

fullmod3 <- glm(party ~ medAge + medHouseIncome + medFamilyIncome +
 pctUnemp + pctPov + pctHS + pctBach + pctBlack + pctHisp,
 family = binomial (link = probit), data = party113cong)

fullmod4 <- glm(party ~ (medAge + medHouseIncome + medFamilyIncome +
 pctUnemp + pctPov + pctHS + pctBach + pctBlack + pctHisp) ^ 2,
  family = binomial (link = probit), data = party113cong)

reducedmod1 <- step(fullmod1, direction = "backward", trace = FALSE)
reducedmod2 <- step(fullmod2, direction = "backward", trace = FALSE)
reducedmod3 <- step(fullmod3, direction = "backward", trace = FALSE)
reducedmod4 <- step(fullmod4, direction = "backward", trace = FALSE)

reducedmod1$aic
reducedmod2$aic
```

```
reducedmod3$aic
reducedmod4$aic

## Part 4

reducedmod2

cor(predict(reducedmod2, type = "response"),
as.numeric(party113cong$party))

# Classification Table, sensitivity and specificity
pi0 <- 0.5

table(Actual=party113cong$party, Predicted =
as.numeric(fitted(reducedmod2) > pi0))

161 / (161 + 41)
204 / (204 + 30)

# Cross-validated Classification Table

pihatcv <- numeric(nrow(party113cong))

for(i in 1:nrow(party113cong))
pihatcv[i] <- predict(update(reducedmod2, subset=-i),
newdata=party113cong[i,],
type="response")

table(Actual=party113cong$party, Predicted = as.numeric(pihatcv > pi0))

145 / (145 + 57)
192 / (192 + 42)

# ROC Curve
png('roccurve.png')

n <- nrow(party113cong)

pihat <- fitted(reducedmod2)

true.pos <- cumsum(as.numeric(party113cong$party)
[order(pihat, decreasing=TRUE)]-1)

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")
```

```
dev.off()

partyhat <- ifelse(fitted(reducedmod2) >= .5, "D", "R")


## Concordance index

mean(outer(pihat[party113cong$party=="D"],
pihat[party113cong$party== "R"], ">") +
0.5 * outer(pihat[party113cong$party=="D"],
pihat[party113cong$party== "R"], "=="))


## Part 5
smallstates <- summary(party113cong$state)
[summary(party113cong$state) ==1 ]

data2 <- party113cong[party113cong$state  %in%
names(smallstates) == FALSE,]

wealth <- ifelse(data2$medHouseIncome > 52000, "Wealthy", "Non-Wealthy" )
wealth <- as.factor(wealth)
wealth <- relevel(wealth, ref = "Non-Wealthy")
data2 <- cbind(data2, wealth)


# Logistic Regression

mod9 <- glm(party ~ state + wealth, family = binomial, data = data2)
summary(mod9)
-1.099 + c(-1, 1) * 1.96 * sqrt(.2943)
exp(-1.099 + c(-1, 1) * 1.96 * sqrt(.2943))

mod10 <- glm(party ~ wealth, family = binomial, data = data2)
summary(mod10)


# Mantel-Haenzsel Analysis


party.array <- xtabs( ~  wealth +party + state,
data = data2, drop.unused.levels = TRUE)


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