Assignment2

Shuyi Yu

11/11/2019

Question1.1

```
as2data <- read.dta("/Users/Shuyi/Desktop/AS2code/conf06.dta")
conf06 <- subset(as2data, as2data$nominee!="ALITO")
vars <- c("vote", "nominee", "sameprty", "qual", "lackqual", "EuclDist2", "strngprs")
conf <- conf06[vars]
conf$numvote <- as.numeric(conf$vote)-1 # from 1/2 (2-yes 1-no) to 0/1 conf
conf$numstrngprs <- as.numeric(conf$strngprs)-1 # same as above
set.seed(123)
samples <- sample(1:nrow(conf), nrow(conf)*0.8, replace = FALSE)
train <- conf[samples, ]
test <- conf[-samples, ]
```

Question1.2

The confusion matrix shows that logit classifier predicts 41 true negatives, 637 true positives, 18 false negatives, and 46 false positives. The overall accuracy is 0.9137466, which is good.

```
logit <- glm(numvote ~ sameprty + qual + EuclDist2 + numstrngprs, data = train, family = binomial)
logit.probs <- predict(logit, newdata = test, type="response")
logit.pred <- ifelse(logit.probs > 0.5, 1, 0)
table(logit.pred, test$numvote)
```

```
##
## logit.pred   0   1
##          0  41  18
##          1  46 637
```

```
mean(logit.pred == test$numvote)
```

```
## [1] 0.9137466
```

Question1.3

The confusion matrix shows that lda classifier predicts 43 true negatives, 630 true positives, 25 false negatives, and 44 false positives. The overall accuracy is 0.9070081, which is slightly lower than that of logit classifier, but close.

```
lda <- lda(numvote ~ sameprty + qual + EuclDist2 + numstrngprs, data = train)
lda.pred <- predict(lda, newdata=test)
table(lda.pred$class, test$numvote)
```

```
##
##       0   1
##   0  43  25
##   1  44 630
```

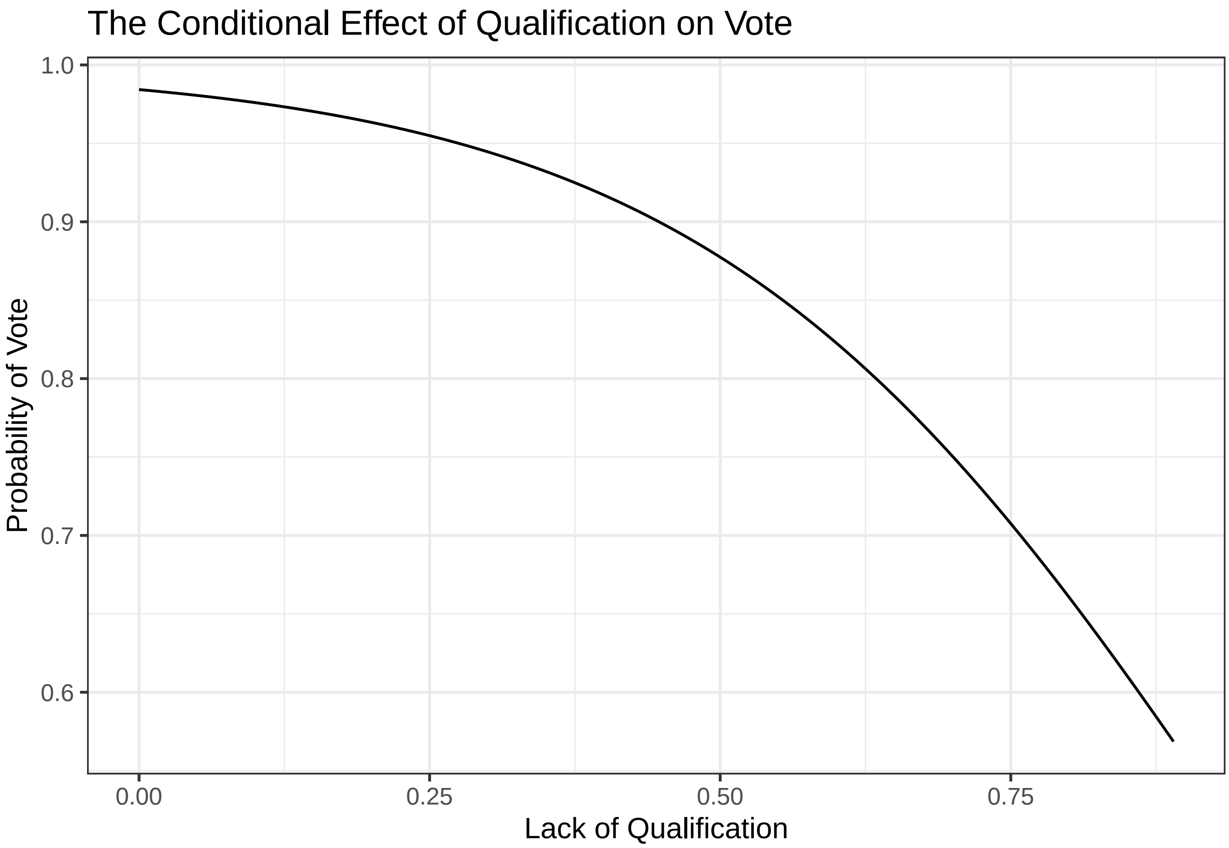
```
mean(lda.pred$class == test$numvote)
```

```
## [1] 0.9070081
```

Question1.4

The plot shows that as the nominees' qualification lessens, the probability of a yes vote decreases. And the decreasing rate picks up as the qualification reaches a lower level, which means that the lack of qualification would pose a greater problem at a lower level.

```r
logit2 <- glm(numvote ~ sameprty + lackqual + EuclDist2 + numstrngprs, data = conf, family = binomial)
conf2 <- with(conf, data.frame(lackqual =
                                 seq(from = min(conf[,5]), to = max(conf[,5]), length.out = 3709),
                                 sameprty = mean(sameprty),
                                 EuclDist2 = mean(EuclDist2),
                                 numstrngprs = mean(numstrngprs)))
conf3 <- cbind(conf2, predict = predict(logit2, newdata = conf2, type = "response"))
ggplot(conf3, aes(x = lackqual, y = predict)) +
  geom_line() +
  labs(x = "Lack of Qualification",
       y = "Probability of Vote") +
  ggtitle("The Conditional Effect of Qualification on Vote") +
  theme_bw()
```



Question1.5

Let's do a summary of the logistic classifier (focus on the z value or p value).

For the variables: (1) the senator sharing the president's party affiliation (sameprty) has a significant positive effect on the probability of a yes vote; (2) the president being strong (numstrngprs) has a significant positive effect on the probability of a yes vote (3) the squared distance between the senator's ideal point and the nominee's inferred ideal point (EuclDist2) has a significant negative effect on the probability of a yes vote. The effects of these three variables indicate that the nomination process is politicized to some extent.

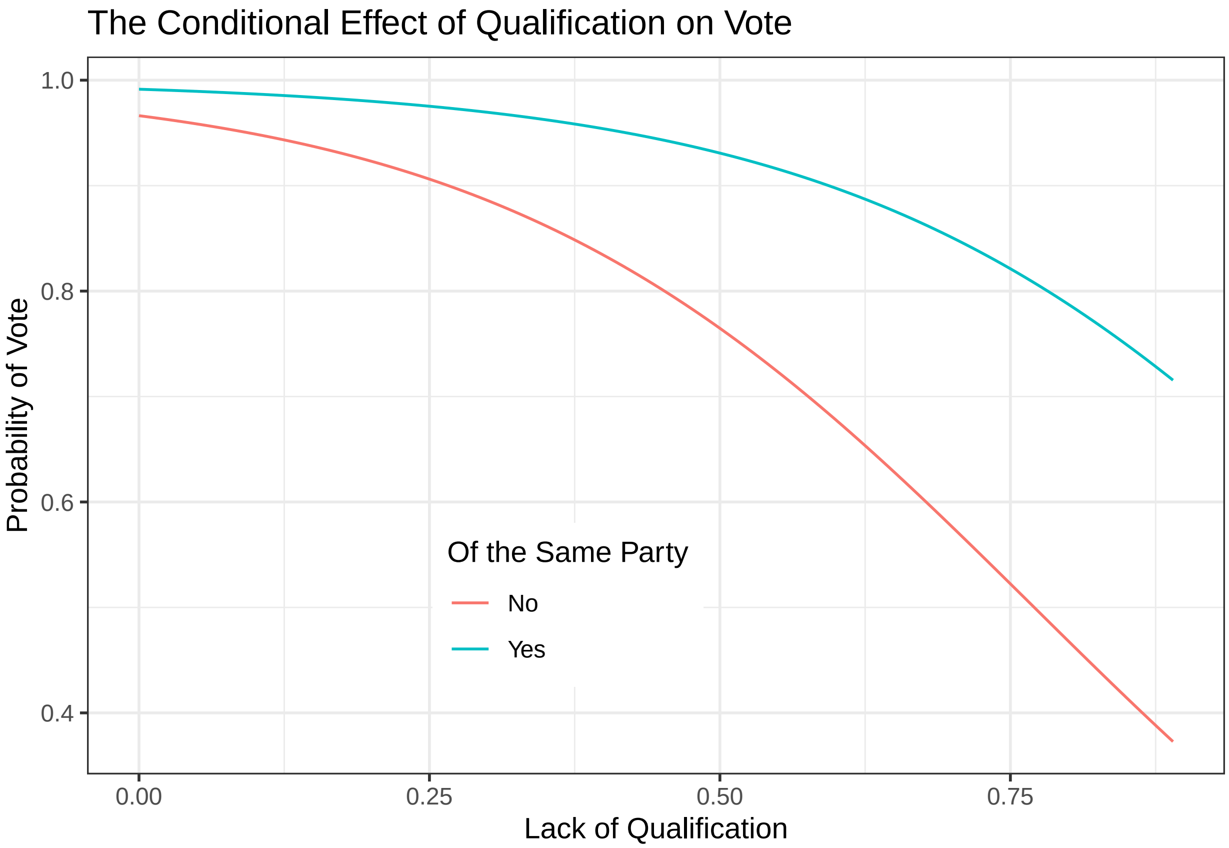
For the variable: the perceived qualification (qual) of the nominee has a significant positive effect on the probability of a yes vote. This is the criteria that an un-politicized nomination process should look at, which has indeed the largest effect, indicating that the nomination process still puts meritocracy in first place.

```r
summary(logit)
```

```
##
## Call:
## glm(formula = numvote ~ sameprty + qual + EuclDist2 + numstrngprs,
##     family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2596   0.0855   0.1955   0.4037   2.1861
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1916     0.2144  -5.558 2.73e-08 ***
## sameprty      1.4673     0.1724   8.511  < 2e-16 ***
## qual          4.4255     0.2645  16.732  < 2e-16 ***
## EuclDist2    -4.0681     0.3213 -12.662  < 2e-16 ***
## numstrngprs   1.5010     0.1524   9.851  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2181.2  on 2966  degrees of freedom
## Residual deviance: 1356.1  on 2962  degrees of freedom
## AIC: 1366.1
##
## Number of Fisher Scoring iterations: 6
```

Question1.6

```
logit3 <- glm(numvote ~ sameprty + lackqual + EuclDist2 + numstrngprs + lackqual*sameprty,
              data = conf, family = binomial)
conf4 <- with(conf, data.frame(lackqual =
                                 rep(seq(from=min(conf[,5]), to=max(conf[,5]), length.out=1854), 2),
                               sameprty = rep(0:1, each = 1854),
                               sameprty = mean(sameprty),
                               EuclDist2 = mean(EuclDist2),
                               numstrngprs = mean(numstrngprs)))
conf5 <- cbind(conf4, predict = predict(logit3, newdata = conf4, type = "response"))
conf5$sameprty <- factor(conf5$sameprty, labels=c("No", "Yes"))
ggplot(conf5, aes(x = lackqual, y = predict, color = sameprty)) +
  geom_line() +
  labs(x = "Lack of Qualification",
       y = "Probability of Vote",
       color = "Of the Same Party") +
  scale_fill_hue(breaks = c("No", "Yes"),
                 labels = c("No",  "Yes")) +
  ggtitle("The Conditional Effect of Qualification on Vote") +
  theme_bw() +
  theme(legend.justification = c(.7,1),
        legend.position = c(.47,.35))
```



Question2.1

I run the algorithm with two dimensions: ideology and race. Two parties are perfectly separated in the two-dimension space. The Democratic party is more liberal (<0) on the first dimension. The Republican party is more conservative (>0) on the first dimension. In general, the Democratic party is more dispersed on the first dimension than the Republican party. And both parties have some variations on the second dimension.

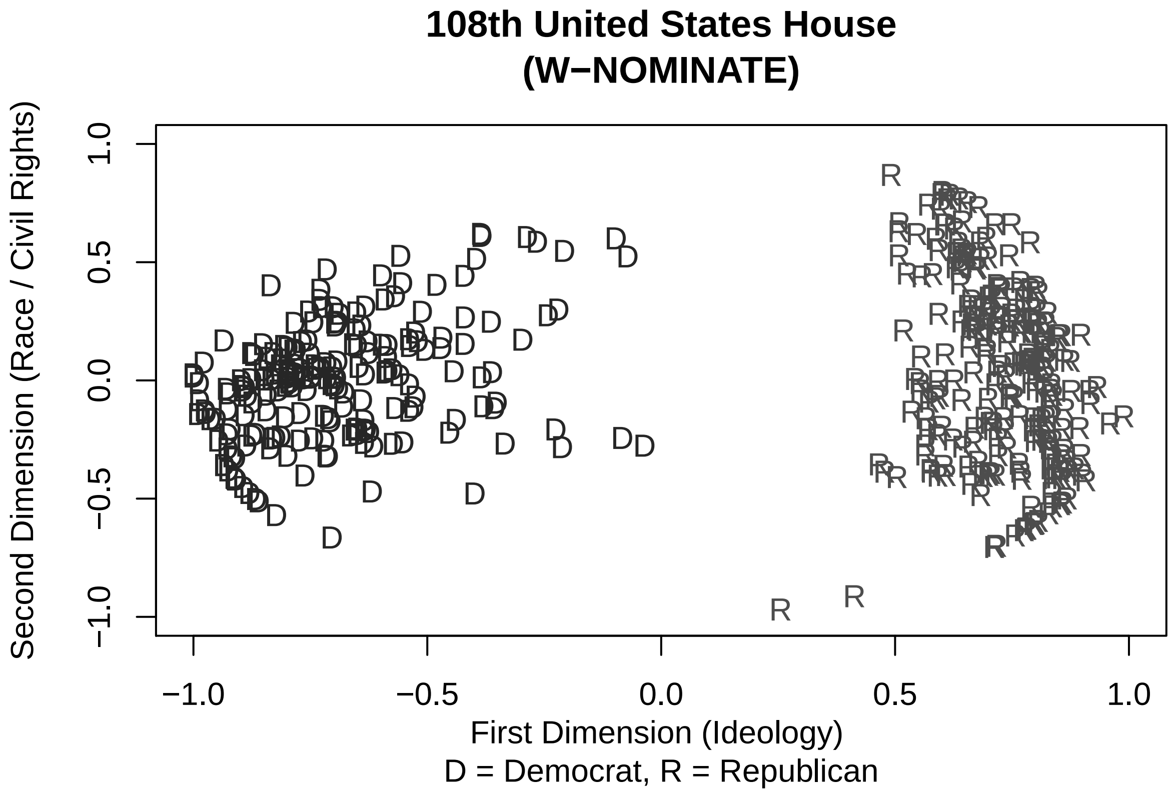
```
house113 <- readKH(
  "/Users/Shuyi/Desktop/AS2code/hou113kh.ord",
  dtl=NULL,
  yea=c(1,2,3),
  nay=c(4,5,6),
  missing=c(7,8,9),
  notInLegis=0,
  desc="113th_House_Roll_Call_Data",
  debug=FALSE
)
```

```
## Attempting to read file in Keith Poole/Howard Rosenthal (KH) format.
## Attempting to create roll call object
## 113th_House_Roll_Call_Data
## 445 legislators and 1202 roll calls
## Frequency counts for vote types:
## rollCallMatrix
##      0      1      6      7      9
##  14576 295753 202943    290  21328
```

```
wnom_result <- wnominate(house113, dims = 2, minvotes = 20, lop = 0.025, polarity = c(2,2))
```

```
##
## Preparing to run W-NOMINATE...
##
##  Checking data...
##
##      ... 1 of 445 total members dropped.
##
##      Votes dropped:
##      ... 181 of 1202 total votes dropped.
##
##  Running W-NOMINATE...
##
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Starting estimation of Beta...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Starting estimation of Beta...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##      Estimating weights...
##      Getting bill parameters...
##      Getting legislator coordinates...
##
##
## W-NOMINATE estimation completed successfully.
## W-NOMINATE took 227.555 seconds to execute.
```

```
par(mfrow = c(1,1))
wnom1 <- wnom_result$legislators$coord1D
wnom2 <- wnom_result$legislators$coord2D
party <- house113$legis.data$party
plot(wnom1, wnom2,
     main="108th United States House\n(W-NOMINATE)",
     xlab="First Dimension (Ideology) \nD = Democrat, R = Republican",
     ylab="Second Dimension (Race / Civil Rights)",
     xlim=c(-1,1), ylim=c(-1,1), type="n")
points(wnom1[party=="D"], wnom2[party=="D"], pch="D", col="gray15")
points(wnom1[party=="R"], wnom2[party=="R"], pch="R", col="gray30")
```



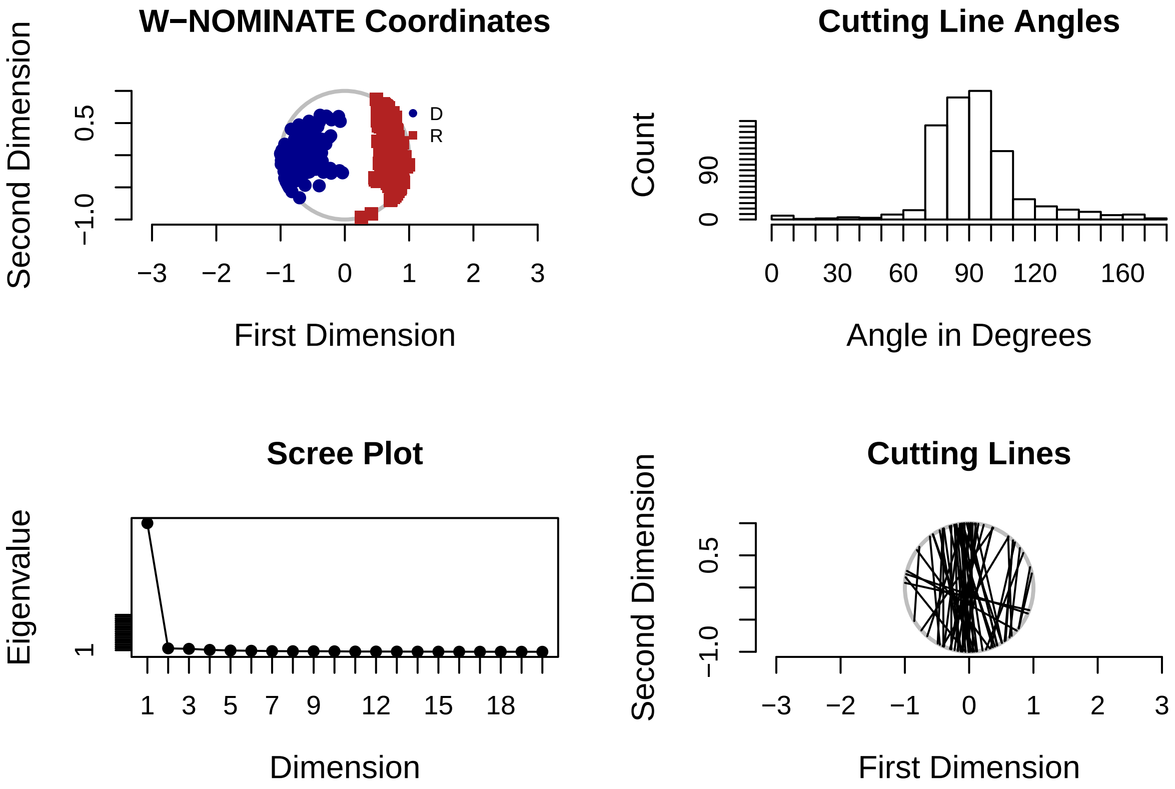
Question2.2

The scree plot shows that the first dimension explains most of the variations. The second and possible subsequent dimensions have much lower explanatory power.

Then look at Correct Classification, APRE and GMP (the first value is for the one-dimensional result, the second value is for the two-dimensional result). The statistics indicate that there is unidimensional structure, with 92.79% of votes correctly classified with a single, ideological dimension. The addition of a second dimension provides only a minimal improvement in fit (Correct Classification from 92.79% to 93.6%, APRE from 92.79% to 93.6%, GMP from 92.79% to 93.6%).

However, following the practice in the class, I would keep the second dimension because it's substantively interesting (race/civil rights).

```
plot(wnom_result)
```



```
## NULL
```

```
summary(wnom_result)
```

```
##
##
## SUMMARY OF W-NOMINATE OBJECT
## ----------------------------
##
## Number of Legislators:     444 (1 legislators deleted)
## Number of Votes:   1021 (181 votes deleted)
## Number of Dimensions:      2
## Predicted Yeas:        212927 of 225718 (94.3%) predictions correct
## Predicted Nays:        185010 of 199413 (92.8%) predictions correct
## Correct Classifiction:     92.79% 93.6%
## APRE:              0.817 0.837
## GMP:           0.84 0.857
##
##
## The first 10 legislator estimates are:
##                   coord1D coord2D
## OBAMA (D USA)      -0.936   0.171
## BONNER (R AL-1)     0.642   0.556
## BYRNE (R AL-1)      0.811   0.205
## ROBY (R AL-2)       0.636   0.772
## ROGERS (R AL-3)     0.724   0.393
## ADERHOLT (R AL-4)   0.678   0.735
## BROOKS (R AL-5)     0.792  -0.007
## BACHUS (R AL-6)     0.632   0.541
## SEWELL (D AL-7)    -0.560   0.024
## YOUNG (R AK-1)      0.565  -0.311
```

Question2.3

Parametric (NOMINATE) and nonparametric (OC) methods for the analysis of preferential choice data can be understood as a trade-off between making strong parametric assumptions about the data and precise estimation of the parameters. In one dimension the OC result is identified only up to a rank order and in two dimensions legislators are identified only to a polytope (i.e. in two dimensions, a legislator could be anywhere inside the specified polytope, but by default OC places him in the center). NOMINATE can recover more detailed metric information about legislator and roll-call characteristics. However, parametric assumptions (e.g., the assumption that errors are iid) can be quite costly. In the cases that these assumptions are violated, OC provides a more accurate picture of preferential choice behavior.