lab1_samir

Samir Datta

September 25, 2017

```r
library(ggplot2)
theme_set(theme_bw())
library(car)
publicopinion <- read.csv('C:/Users/Samir/Documents/MIDS/StatsF17/lab 1/public_opinion.csv')

publicopinion$partyfactor <- ifelse(publicopinion$party==1, 'Democrat',
                                    ifelse(publicopinion$party==2, 'Other',
                                           'Republican'))

publicopinion$age <- 2017 - publicopinion$birthyr
publicopinion$genderfactor <- ifelse(publicopinion$gender==1,'Male', 'Female')
publicopinion$racefactor <- ifelse(publicopinion$race_white==1, 'White', 'Non-White')
publicopinion$spfactor <- ifelse(publicopinion$sanders_preference==1, "Yes", "No")
publicopinion_narm <- publicopinion[!is.na(publicopinion$sanders_preference),]
```

Question 1: Model the relationship between age and voters' preference for Bernie Sanders over Hillary Clinton. Select the model that you prefer the most and describe why you chose these variables and functional form.

Question 1a: Describe your chosen model in words, along with a brief description of the variables and the model's functional form (Note: You do not have to justify your choices at this step).

```r
glm.out2 <- glm.out <- glm(sanders_preference ~ age*genderfactor+partyfactor+racefactor,
               data=publicopinion_narm,
               family=binomial(link="logit"))

anova(glm.out, glm.out2, test="LR")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age * genderfactor + partyfactor + racefactor
## Model 2: sanders_preference ~ age * genderfactor + partyfactor + racefactor
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1184     1529.7
## 2      1184     1529.7  0        0
```

```r
glm.out <- glm(sanders_preference ~ age*racefactor+partyfactor+racefactor,
               data=publicopinion_narm,
               family=binomial(link="logit"))
summary(glm.out)
```

```
##
```

```
## Call:
## glm(formula = sanders_preference ~ age * racefactor + partyfactor +
##     racefactor, family = binomial(link = "logit"), data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6715  -1.1874   0.8050   0.9765   1.7927
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.236404   0.356396   0.663 0.507126
## age                   -0.020500   0.007634  -2.685 0.007243 **
## racefactorWhite        0.397963   0.417595   0.953 0.340597
## partyfactorOther       0.708699   0.140518   5.043 4.57e-07 ***
## partyfactorRepublican  0.586729   0.163120   3.597 0.000322 ***
## age:racefactorWhite    0.010485   0.008702   1.205 0.228266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1531.7  on 1185  degrees of freedom
## AIC: 1543.7
##
## Number of Fisher Scoring iterations: 4
```

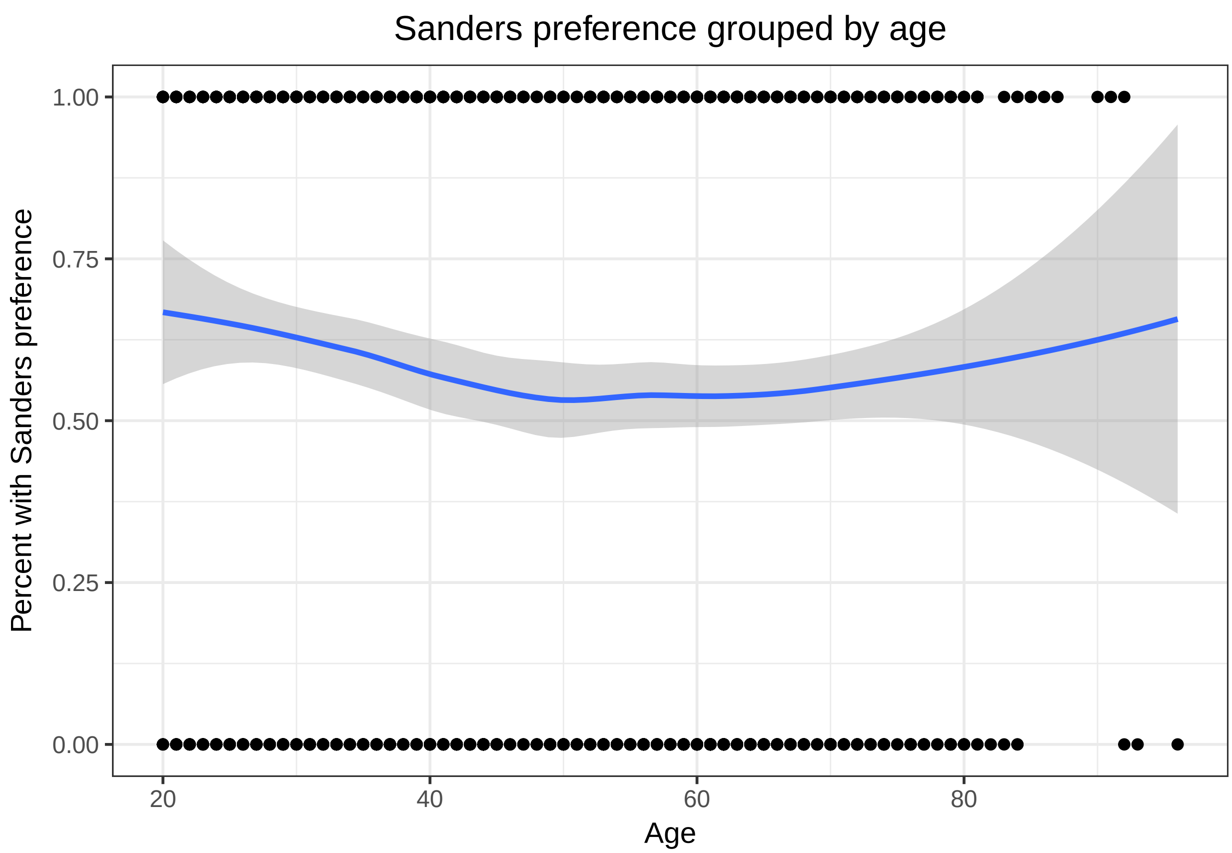
We chose a logistic regression model to predict Sanders preference among voters using the logit link function. As predictors in the model, in addition to the main variable of interest age, we have included political party (Democrat, Republican, or Other) and race (Whte or Non-White)

Question 1b: Describe the variables your have included in your model and justify why you chose these variables and the model's functional form.

Age

```r
ggp <- ggplot(publicopinion_narm, aes(x=age, y=sanders_preference))

ggp + geom_point()+
  geom_smooth(method="loess", se=T)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age")+
  theme(plot.title=element_text(hjust=.5))
```



Above is a scatterplot of all of the points in the set, with age on the x-axis and the binary variable sanders_preference on the y-axis. Displaying the dots like this is not necessarily informative, but the loess smooth curve - and standard error ribbon - reveals an interesting trend. Below the age 50, there seems to be a trend for younger voters to prefer Sanders. However, this trend is also seen in the opposite direction for voters above around 70. This would suggest that including a quadratic term for age might be useful. However, it is important to note that the standard error ribbon is very large towards the older end of the age range, which is indicative of how few voters of that age range we really have. While we should try modeling a quadratic term for age, we should be careful not to over-interpret any result based off insufficient data.

```
length(publicopinion_narm[publicopinion_narm$age>70,1])*100/
  length(publicopinion_narm[,1])
```

```
## [1] 10.99916
```

Only 11% of voters in the sample are older than 70.

Political party

Party is a three-level categorical variable with levels of Democrat, Reupublican, or other.

```
po_party_agg <- with(publicopinion_narm,
                     aggregate(cbind(100*sanders_preference),
                               list(partyfactor=partyfactor),
                               mean))
po_party_agg$n <- with(publicopinion_narm,
                     aggregate(cbind(100*sanders_preference),
                               list(partyfactor=partyfactor),
```

```
                                length))[,2]

po_party_df <- data.frame(sanders_pref_percent=po_party_agg$V1,
                          party=po_party_agg$partyfactor,
                          sample_percent = 100*po_party_agg$n/1191)
po_party_df
```

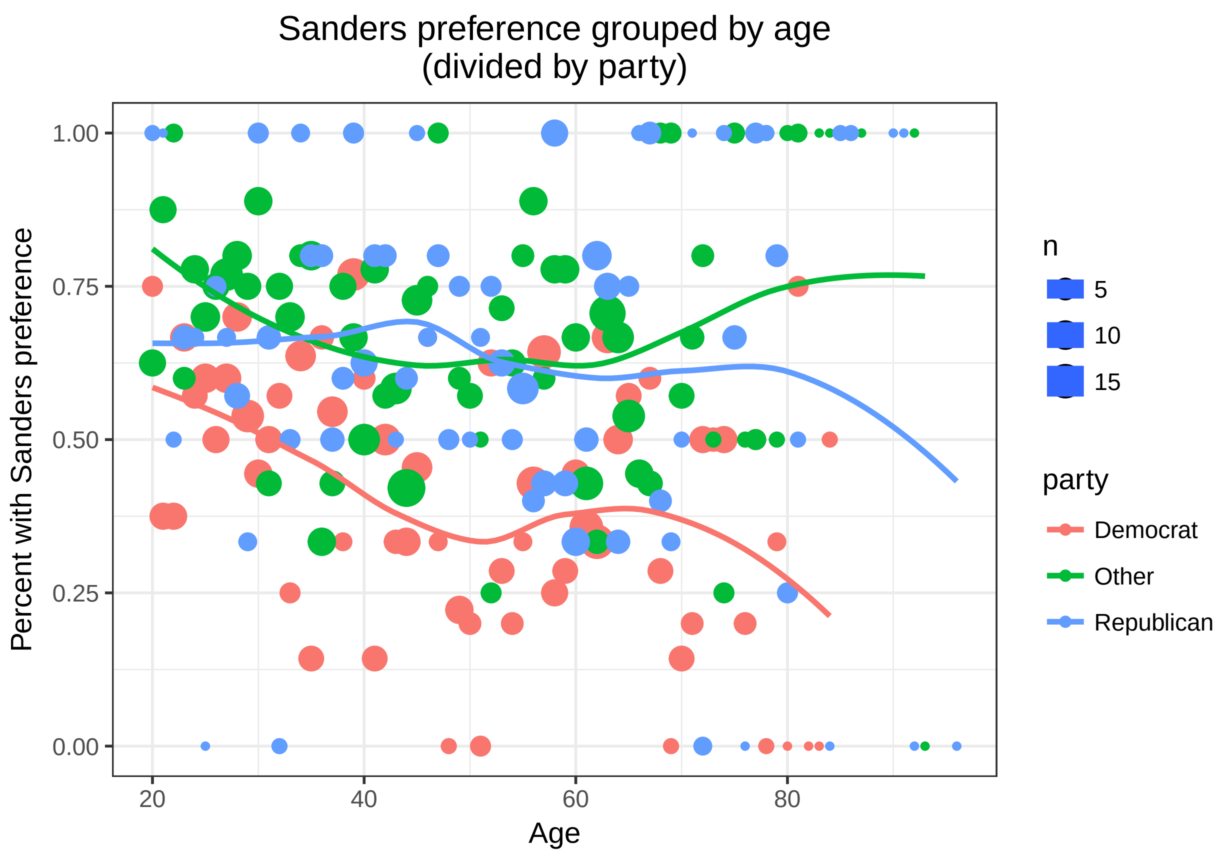
```
##   sanders_pref_percent      party sample_percent
## 1             45.27473   Democrat       38.20319
## 2             65.93886      Other       38.45508
## 3             64.02878 Republican       23.34173
```

Republicans, who make up 23% of the sample, are slightly underrepresented compared to Democrats and Other, but not to the extent that we should be worried about sampling bias. A slight majority (55%) of Democrats polled preferred Clinton to Sanders, while a majority of Republicans (64%) and Other (66%) preferred Sanders. This difference supports including party as an explanatory variable.

```
age_bin_agg_party <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       party=partyfactor), mean))
age_bin_agg_party$n <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       party=partyfactor), length))[,3]

ggp <- ggplot(age_bin_agg_party, aes(x=agebin, y=sanders_preference,
                                     color=party, size=n))

ggp + geom_point(aes(color=party))+
  geom_smooth(method="loess", se=F)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age\n(divided by party)")+
  theme(plot.title=element_text(hjust=.5))
```



Above is a scatterplot of age on the x-axis, where color represents the political party. Each dot's position on the y-axis represents the percent of people of that specific party and age that preferred Sanders.

The democrats seem to have a fairly clear relationship with age in that younger democrats look more likely to support Sanders than older ones. The relationship within Republicans is less clear, and for independents, it looks almost quadratic (as the smooth curve lifts upwards both for younger and older voters). The curves are loess smoothed curves and not meant to be a perfect representation of overall trends. However, there is still enough evidence to support at least trying to model an age by party interaction, since it looks like different parties may have different relationships with age.

Gender

```
po_gender_agg <- with(publicopinion_narm,
                      aggregate(cbind(100*sanders_preference),
                                list(genderfactor=genderfactor),
                                mean))
po_gender_agg$n <- with(publicopinion_narm,
                      aggregate(cbind(100*sanders_preference),
                                list(genderfactor=genderfactor),
                                length))[,2]

po_gender_df <- data.frame(sanders_pref_percent=po_gender_agg$V1,
                           party=po_gender_agg$genderfactor,
                           sample_percent = 100*po_gender_agg$n/1191)
po_gender_df
```

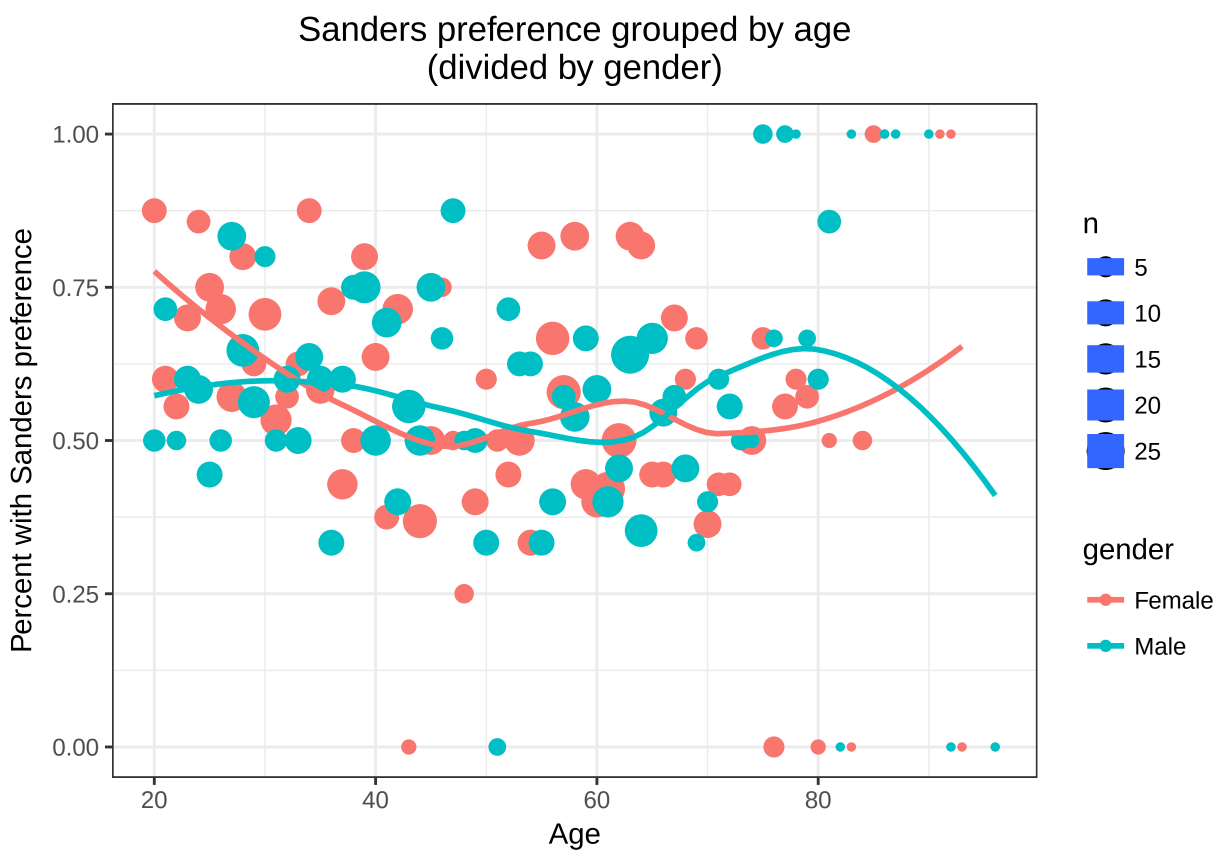
```
##   sanders_pref_percent  party sample_percent
## 1             57.71704 Female       52.22502
## 2             57.46924   Male       47.77498
```

Male and female voters are close to equally represented, both around 50%. Across the sample male and female voters prefer Sanders at almost the same rate (57.5% for male, 57.7% for female)

```
age_bin_agg_gender <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       gender=genderfactor), mean))
age_bin_agg_gender$n <- with(publicopinion_narm,
                        aggregate(cbind(sanders_preference),
                                  list(agebin=age,
                                       gender=genderfactor), length))[,3]

ggp <- ggplot(age_bin_agg_gender, aes(x=agebin, y=sanders_preference,
                                      color=gender, size=n))

ggp + geom_point(aes(color=gender))+
  geom_smooth(method="loess", se=F)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age\n(divided by gender)")+
  theme(plot.title=element_text(hjust=.5))
```



Towards the younger end of the age range, it seems like the negative relationship between sanders preference and age exists mostly for female voters and not so much for male voters. This suggests that investigating a gender by age interaction may be useful.

Race

```
po_race_agg <- with(publicopinion_narm,
                    aggregate(cbind(100*sanders_preference),
                              list(racefactor=racefactor),
                              mean))
po_race_agg$n <- with(publicopinion_narm,
                    aggregate(cbind(100*sanders_preference),
                              list(racefactor=racefactor),
                              length))[,2]

po_race_df <- data.frame(sanders_pref_percent=po_race_agg$V1,
                         party=po_race_agg$racefactor,
                         sample_percent = 100*po_race_agg$n/1191)
po_race_df
```

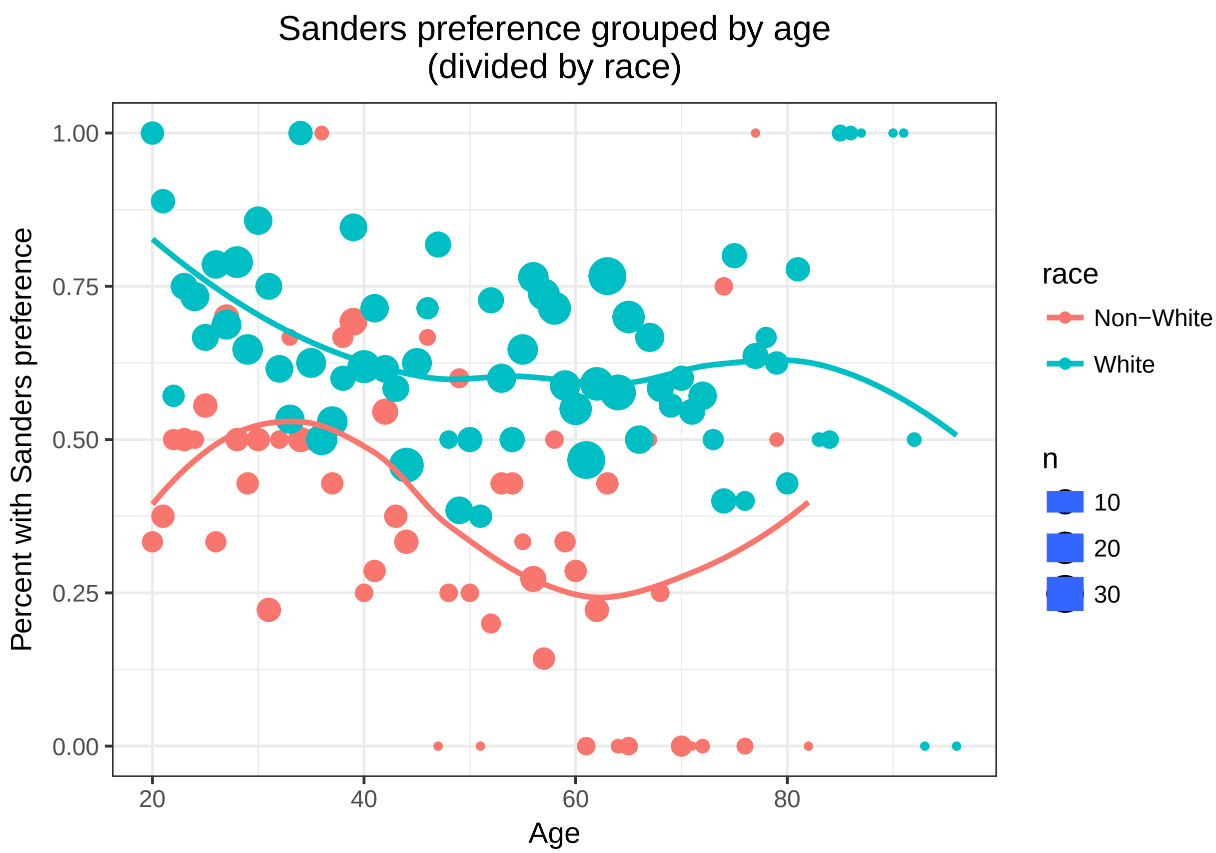
```
##   sanders_pref_percent     party sample_percent
## 1             40.99379 Non-White        27.0361
## 2             63.75144     White        72.9639
```

White voters make up about 73% of the sample, which is close to the estimated percentage of White people in America, which further supports our sample being representative. There is a large difference between how many white voters (64%) vs. non-whie voters (41%) prefer Sanders.

```
age_bin_agg_race <- with(publicopinion_narm,
                       aggregate(cbind(sanders_preference),
                                 list(agebin=age,
                                      race=racefactor), mean))
age_bin_agg_race$n <- with(publicopinion_narm,
                       aggregate(cbind(sanders_preference),
                                 list(agebin=age,
                                      race=racefactor), length))[,3]

ggp <- ggplot(age_bin_agg_race, aes(x=agebin, y=sanders_preference,
                                     color=race, size=n))

ggp + geom_point(aes(color=race))+
  geom_smooth(method="loess", se=F)+
  ylab("Percent with Sanders preference")+
  xlab("Age")+
  ggtitle("Sanders preference grouped by age\n(divided by race)")+
  theme(plot.title=element_text(hjust=.5))
```



At first glance the relationship seems very different for non-white voters, but this may be an artifact of the loess smoothing curve attempting to compensate for the data points in the youngest age groups. When looking at the overall distribution of the dots it seems that both white and non-white voters have a negative relationship with age.

Question 1c: Based on your EDA, describe other models that you might have considered and why you ended up choosing your ???nal model. Be sure to print each of the model results and any statistical tests you used to choose which model to use.

In our EDA, we determined that race and party had large effects on Sanders preference and would be important to control for. Gender did not seem like it had much explanatory power on its own, although a gender by age interaction seemed plausible. A race by age interacton also looked to be worth testing. Finally, we wanted to test the plausability of a quadratic age term.

Gender by age interaction

```
glm.out.base <- glm(sanders_preference ~ age + partyfactor + racefactor +
                  genderfactor, data=publicopinion_narm,
                  family=binomial(link='logit'))
glm.out.int <- glm(sanders_preference ~ age + partyfactor + racefactor +
                  genderfactor + age:genderfactor, data=publicopinion_narm,
                  family=binomial(link='logit'))
```

```r
summary(glm.out.base)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor, family = binomial(link = "logit"), data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7263  -1.1765   0.7857   0.9837   1.7032
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.058622   0.212382  -0.276 0.782531
## age                   -0.012602   0.003671  -3.433 0.000598 ***
## partyfactorOther       0.731136   0.141515   5.166 2.39e-07 ***
## partyfactorRepublican  0.601001   0.163208   3.682 0.000231 ***
## racefactorWhite        0.877155   0.142044   6.175 6.61e-10 ***
## genderfactorMale      -0.129182   0.123222  -1.048 0.294472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1532.1  on 1185  degrees of freedom
## AIC: 1544.1
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(glm.out.int)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor + age:genderfactor, family = binomial(link = "logit"),
##     data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7841  -1.1721   0.8037   0.9526   1.6738
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.193790   0.269511   0.719 0.472114
## age                   -0.017725   0.004981  -3.559 0.000373 ***
## partyfactorOther       0.731612   0.141670   5.164 2.41e-07 ***
## partyfactorRepublican  0.604385   0.163439   3.698 0.000217 ***
## racefactorWhite        0.881578   0.142283   6.196 5.79e-10 ***
## genderfactorMale      -0.675620   0.376790  -1.793 0.072958 .
## age:genderfactorMale   0.011047   0.007195   1.535 0.124710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1529.7  on 1184  degrees of freedom
## AIC: 1543.7
##
## Number of Fisher Scoring iterations: 4
```

Party by age interaction

```r
glm.out.base <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))
glm.out.int <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor + age:partyfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))

summary(glm.out.base)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor, family = binomial(link = "logit"), data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7263  -1.1765   0.7857   0.9837   1.7032
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.058622   0.212382  -0.276 0.782531
## age                  -0.012602   0.003671  -3.433 0.000598 ***
## partyfactorOther      0.731136   0.141515   5.166 2.39e-07 ***
## partyfactorRepublican  0.601001   0.163208   3.682 0.000231 ***
## racefactorWhite       0.877155   0.142044   6.175 6.61e-10 ***
## genderfactorMale     -0.129182   0.123222  -1.048 0.294472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1532.1  on 1185  degrees of freedom
## AIC: 1544.1
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(glm.out.int)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor + age:partyfactor, family = binomial(link = "logit"),
##     data = publicopinion_narm)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6971  -1.1605   0.8035   0.9550   1.7720
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                0.179279   0.300649   0.596  0.55097
## age                       -0.017556   0.005764  -3.046  0.00232 **
## partyfactorOther           0.357773   0.422464   0.847  0.39707
## partyfactorRepublican      0.144621   0.505968   0.286  0.77501
## racefactorWhite            0.878499   0.142220   6.177 6.53e-10 ***
## genderfactorMale          -0.129135   0.123317  -1.047  0.29502
## age:partyfactorOther       0.007719   0.008240   0.937  0.34888
## age:partyfactorRepublican  0.009072   0.009379   0.967  0.33341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1530.8  on 1183  degrees of freedom
## AIC: 1546.8
##
## Number of Fisher Scoring iterations: 4
```

There are a number of reasons that modeling an age by party interaction does not seem like a good idea. First of all, neither of the interacton terms (age:Other, age:Republican) have large effects. Their coefficients are relatively small compared to the original coefficient of the age term, and their p-values are nowhere near statistical significance ($p > .33$). The AIC for the model with the interaction term is larger than for the model without it, suggesting that our additional model complexity is not helping the overall model.

```
anova(glm.out.base, glm.out.int, test="LR")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age + partyfactor + racefactor + genderfactor
## Model 2: sanders_preference ~ age + partyfactor + racefactor + genderfactor +
##     age:partyfactor
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1185     1532.1
## 2      1183     1530.8  2   1.2789   0.5276
```

The likelihood ratio test, with a p-value of .53, also suggests our interaction model is not more useful than the simpler model. For these reasons we decided to not model an age by party interaction.

Quadratic age term

```
glm.out.base <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor, data=publicopinion_narm,
                 family=binomial(link='logit'))
glm.out.quad <- glm(sanders_preference ~ age + partyfactor + racefactor +
                 genderfactor + I(age^2), data=publicopinion_narm,
                 family=binomial(link='logit'))
```

```r
summary(glm.out.base)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor, family = binomial(link = "logit"), data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7263  -1.1765   0.7857   0.9837   1.7032
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.058622   0.212382  -0.276 0.782531
## age                    -0.012602   0.003671  -3.433 0.000598 ***
## partyfactorOther        0.731136   0.141515   5.166 2.39e-07 ***
## partyfactorRepublican  0.601001   0.163208   3.682 0.000231 ***
## racefactorWhite         0.877155   0.142044   6.175 6.61e-10 ***
## genderfactorMale       -0.129182   0.123222  -1.048 0.294472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1532.1  on 1185  degrees of freedom
## AIC: 1544.1
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(glm.out.quad)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor +
##     genderfactor + I(age^2), family = binomial(link = "logit"),
##     data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7853  -1.1679   0.7913   0.9462   1.6336
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.8125517  0.5165613   1.573 0.115718
## age                    -0.0519434  0.0215906  -2.406 0.016136 *
## partyfactorOther        0.7353441  0.1418181   5.185 2.16e-07 ***
## partyfactorRepublican  0.6031312  0.1633682   3.692 0.000223 ***
## racefactorWhite         0.8722500  0.1425814   6.118 9.50e-10 ***
## genderfactorMale       -0.1209202  0.1234752  -0.979 0.327428
## I(age^2)                0.0003921  0.0002120   1.849 0.064446 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1528.6  on 1184  degrees of freedom
## AIC: 1542.6
##
## Number of Fisher Scoring iterations: 4
```

```
anova(glm.out.base, glm.out.quad, test="LR")
```

```
## Analysis of Deviance Table
##
## Model 1: sanders_preference ~ age + partyfactor + racefactor + genderfactor
## Model 2: sanders_preference ~ age + partyfactor + racefactor + genderfactor +
##     I(age^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1185     1532.1
## 2      1184     1528.6  1   3.4751   0.0623 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Like the EDA showed, the model is showing that a quadratic age term might be plausible. The significance of the term in the model is slightly larger than .05 (.06), as is the signifiance of the likelihood ratio test (.06). The AIC of the model with the quadratic term is also lower.

However, we decided to not to include the quadratic age term in the end. Aside from the lack of statistical significance - although it is close - the main reason for this is the lack of representation in the older age range that is driving this result. We would not feel comfortable recommending this model and suggesting that older voters be targeted when we have so few older people that are contributing to this trend.

```
glm.out <- glm(sanders_preference ~ age+partyfactor+racefactor,
               data=publicopinion_narm,
               family=binomial(link="logit"))
summary(glm.out)
```

```
##
## Call:
## glm(formula = sanders_preference ~ age + partyfactor + racefactor,
##     family = binomial(link = "logit"), data = publicopinion_narm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7036  -1.1792   0.7907   0.9881   1.6662
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.115017   0.205360  -0.560 0.575428
## age                  -0.012480   0.003666  -3.404 0.000664 ***
## partyfactorOther      0.713501   0.140368   5.083 3.71e-07 ***
## partyfactorRepublican  0.594231   0.162972   3.646 0.000266 ***
## racefactorWhite        0.872782   0.141872   6.152 7.66e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1623.5  on 1190  degrees of freedom
## Residual deviance: 1533.2  on 1186  degrees of freedom
## AIC: 1543.2
##
## Number of Fisher Scoring iterations: 4
```

```
glm.out$coefficients
```

```
##           (Intercept)                   age      partyfactorOther
##           -0.11501737           -0.01248024            0.71350064
## partyfactorRepublican       racefactorWhite
##            0.59423118            0.87278200
```

```
a = c(20:100)
y = exp(glm.out$coefficients[1] + glm.out$coefficients[2]*a)/
       (1+exp(glm.out$coefficients[1] + glm.out$coefficients[2]*a))
```

```
plot(a, y)
```